An Evaluation of the Merging Interaction between Humans and Interaction-Aware Vehicles

By

Federico Scarì

in partial fulfilment of the requirements for the degree of

Master of Science

in Mechanical Engineering (Vehicle Engineering)

at the Delft University of Technology,
to be defended publicly on Tuesday March 26, 2024 at 13:00

Student Number:	5377013	
Thesis committee:	Ir. O. Siebinga,	TU Delft, Supervisor
	Dr. A. Zgonnikov,	TU Delft, Supervisor
	Prof. dr. ir. D.A. Abbink,	TU Delft, External member

An electronic version of this thesis is available at http://repository.tudelft.nl/.
The codes are at https://github.com/fscari/varjo.git and https://github.com/fscari/driving-interactions.git
The dataset is available at 4TU.ResearchData https://data.4tu.nl

An Evaluation of the Merging Interaction between Humans and Interaction-Aware Vehicles

Federico Scarì
Department of Cognitive Robotics, Delft University of Technology, Faculty of 3mE

Abstract—As autonomous vehicle (AV) technology progresses, the necessity for a comprehensive understanding of interactions between AVs and human-driven vehicles (HVs) becomes paramount, particularly in critical manoeuvres such as merging. Mastering merging interactions is essential for enhancing road safety. Existing research in this field focuses on how the AV performs the merging manoeuvre but often fails to assess how they influence these interactions. By drawing inspiration from Human-Robot Interaction and Human Aware Navigation, this study aims to bridge this gap by examining how these interactions influence driver workload, measured through fixations duration, perceived safety and drivers' subjective perception during merging scenarios. We employed a Virtual Reality environment to simulate realistic driving conditions and measure driver responses. We conducted an experiment where participants engaged in merging manoeuvres with each other and, subsequently and without being informed, with the AV described in "Planning for cars that coordinate with people" [1]. This approach allowed for an unbiased assessment of natural driver reactions to AV behaviours. Our findings reveal significant increases in driver workload and decreases in perceived safety during HV-AV interactions, compared to HV-HV interactions. These results suggest that current AV algorithms may not fully account for the complexity of human-AV interactions, highlighting a need for interaction evaluation in the AV development. Participants' subjective feedback indicates a recognition of and negative reaction to AV driving behaviours, emphasizing the importance of designing AVs that are both efficient and intuitive for human drivers. The study's implications suggest improving AV controllers' evaluations by including their interactions with human drivers. By integrating interaction evaluation, AV technologies can achieve smoother and more successful integration into existing road systems, enhancing predictability and driver acceptance. This study marks a step towards understanding the interactions between AVs and HVs, offering insights that could steer future research and development in autonomous driving technologies.

Index Terms— Human-Human Interaction, Human-AV Interaction, Workload, Perceived Safety

1 Introduction

Autonomous vehicles (AV) are recognized as the future of transportation, with experts discussing their advantages and disadvantages [2, 3]. While consensus on the timeline and implementation of fully autonomous vehicles is lacking, a gradual increase in automation levels is anticipated, resulting in a mixed traffic flow with manually, semi-autonomously, and fully autonomously driven cars [4]. This shift will transform driving habits, impacting safety, efficiency, and traffic flow. The integration of autonomous vehicles into the existing road system relies on acceptance by human drivers in human-driven vehicles (HV). The concept of "acceptance" in the context of AVs is an important term that plays a crucial role in the interaction between AVs and HVs [5]. Acceptance encompasses a variety of intricate concepts like perceived ease of use, attitude, social norm, trust, perceived usefulness, perceived risk, compatibility [5] as well as the predictability of AV behaviour, also called legibility [6]. Understanding and enhancing this acceptance is crucial, as it directly impacts the effectiveness of AV integration into our transportation systems and influences the overall safety and efficiency of road traffic [5, 7]. The importance of perceived safety in human-AV interactions [5], in influencing HV-AV interactions, is a key aspect that researchers must consider, similarly to the safety of human-robot interactions in broader contexts [8].

However, there is a notable gap in the current field of research about the interaction between humans and AVs [9]. Recent advancements in AV technology in this field have led to the development of interaction-aware controllers (IACs) [10]. These IACs integrate a predictive model of human driver behavior into their systems, enabling them to predict potential responses from human drivers to the AV's manoeuvres. Leveraging these predictions and a reward function that prioritizes aspects like safety and comfort, the IAC identifies the most appropriate action for the AV. Many studies that focus on AV controllers and IACs for merging manoeuvres [1, 11–19], claiming to enhance AV-HV interactions, lack empirical validation in real-world settings or controlled simulations with human participants that represent these reality-like conditions, and thus overlooking the dynamics of road interactions. This raises concerns about the validity of their generalizability to actual traffic scenarios, as showcasing a proposed controller in a simulated traffic setting does not adequately indicate its effectiveness in real-world conditions. In fact, assessments of an AV controller for merging (lane change and intersections) focus on vehicle speed and acceleration profiles [1, 11, 17], execution time [12, 17], safety margins [1, 15, 16], and velocity adjustments to minimize a cost function [17], with evaluations primarily conducted through simulations without involving human drivers. This approach overlooks real-world interactions with human drivers, failing to account for the AV's impact on human behaviour and their mutual interaction. However, understanding and smoothing out these interactions is crucial, as they directly impact road safety, traffic efficiency, and the overall user experience for both AV users and HV drivers. The absence of effective communication between AVs and HVs can lead to misunderstandings or misinterpretations of intent, potentially causing hazardous situations.

Hence, the challenge lies in creating and evaluating AV systems that not only excel in autonomous navigation but also in 'social' driving skills. These skills encompass understanding and anticipating human driver behaviours, communicating intentions implicitly or explicitly, and making decisions that are predictable and comprehensible to human drivers. Not only designing but being able to evaluate these aspects is paramount to ensure seamless and safe integration of AVs into our road systems, particularly in (highways) merging scenarios where high speeds, complex driving patterns and their impact on road safety (lane-changing and merging manoeuvres contribute to over 450000 accidents in the United States alone (2015) [16]) necessitate a deeper level of coordination and cooperation [20] between all vehicles on the road. In fact, tasks such as lane changing and merging fall within the first and second levels of Michon's model [21], the most widely used framework to explain driving behaviour, and play a crucial role in driving outcomes. Especially the tactical level directly influences the interactions between vehicles, impacting the overall high-level success of merging manoeuvres. In summary, current research overlooks interactions between autonomous and human-driven vehicles, neglecting crucial elements like human acceptance, human perceived safety, and anticipatory behaviour. An alternative way to evaluate AV is to address the interaction of AV and HV by evaluating the impact of AVs on HVs. There are fields like HRI and HAN that already address the interaction between humans and robots. Drawing inspiration from these fields, this work aims to evaluate the impact of the controller proposed in [1] on an HV based on two key assessment criteria. These criteria – workload and perceived safety – are chosen for their relevance in assessing not only the technical efficacy but also the effectiveness of passive communication and interaction between autonomous vehicles and human drivers. In addition to evaluating perceived safety, comparing workload levels between human-human (HV-HV) and human-autonomous vehicle (HV-AV) interactions is vital. While autonomous vehicles aim to reduce the workload of drivers, they should not inadvertently lead to an increased workload for human drivers.

1.1 This Study

This work aims to answer the following research questions:

1) How does the merging behaviour of interaction-aware controllers for autonomous vehicles compare to that of human drivers influence the high-level merging outcomes, as measured by:
   a) the number of collisions, and
   b) the percentage of merges in front of the other vehicle?
2) How does the merging interaction between interaction-aware controllers for autonomous vehicles and human-driven vehicles compare to interactions between two human-driven vehicles, in terms of:
   a) the human driver's workload, and
   b) the human driver's perceived safety?
3) How do human drivers perceive and distinguish the driving behaviours of autonomous vehicles compared to those of human-driven vehicles in a merging manoeuvre?

To answer these research questions, we conducted a driving experiment where two participants had to interact with each other in a merging scenario. The experiment took place in a Virtual Reality (VR) environment, where eye movements were recorded and used to measure the driver's workload. Furthermore, although the participants were told that they would interact with each other, in half of the experiment, they interacted with AVs instead of with other participants. This approach was adopted to avoid any possible bias that participants could have towards AVs. During these experiments, the AV's impact on the human drivers during the merging manoeuvre was evaluated through the participants' workload, perceived safety, and questionnaires. When simplifying a two-vehicle merging manoeuvre, the scenario can be conceptualized as involving two distinct agents: one merging onto the highway and the other already cruising on it. In this context, the study focuses on analyzing the participants' perceived safety and workload in both cruising and merging roles, to find any differences between human-human (HV-HV) and human-autonomous vehicle (HV-AV) interactions.

Our initial hypotheses are the following:

1) *In the context of merging manoeuvres, the high-level merging outcomes between autonomous vehicles with interaction-aware controllers and human-driven vehicles, as compared to interactions between two human-driven vehicles, will result in:*
   a) *similar number of collisions*
   b) *similar percentage of merges in front of the other vehicle*
2) *In the context of merging manoeuvres, the interaction between autonomous vehicles with interaction-aware controllers and human-driven vehicles, as compared to interactions between two human-driven vehicles, will result in:*
   a) *an increased workload for the human driver*
   b) *a decreased perceived safety for the human driver*
3) *Human drivers will be able to* ***recognize differences in driving behaviour between autonomous vehicles and human-driven vehicles*** *and will perceive the AV driving behaviour differences as unsettling, bothersome or dangerous*

The idea behind the first hypothesis relies on the fact that such an AV controller is already evaluated based on similar technical parameters and thus optimized for such evaluation. Nevertheless, the rationale behind the other hypotheses is grounded in the behavioural patterns and decision-making processes of AVs, which could be unfamiliar and potentially unsettling for human drivers [22]. Furthermore, human drivers are likely to recognize and react negatively to these behavioural differences [23]. This study aims to investigate these hypotheses, thereby contributing to the understanding of human-AV interaction dynamics.

2 Example of Interaction-aware controller

The controller proposed in [1] was selected as the controller for this experiment for its on-paper technical capabilities and its potential for seamless interactions with human drivers. Additionally, the selection was influenced by the availability of its code on the author's GitHub repository, making it easily accessible for assessment.

The controller proposed focuses on enhancing the interaction between AVs and HVs. It acknowledges that AV actions influence HV reactions, which can lead to coordination between AVs and HVs. Unlike most AV controllers, which treat HVs as obstacles, this controller is designed to enable AVs to positively impact HV responses, improving efficiency and cooperation.

To achieve this, the AV-HV interaction is framed as a partially observable stochastic game (POSG). However, to let the controller estimate the HV's reward function online, the problem was converted into a partially observable Markov decision process (POMDP).

This approach allows the AV to choose a trajectory and anticipate HV responses by knowing only the HV's initial values. Furthermore, the AV can "probe" the human's reactions through an information gain term in the reward function, optimizing for expected reward and considering the effects of its actions on observations. Active information gathering also involves testing the human's reactions based on the collision avoidance parameter. Model Predictive Control (MPC) is utilized to perform these optimizations. This approach is based on the main concept outlined in the paper, which is to emphasize the "coordination" between the AV and the HV in such scenarios. Although the HV is not merely seen as an obstacle by this AV controller, as the AV is the first to initiate action, it may come across as assertive or even aggressive towards the HV, potentially harming the interaction and leading to increased workload and reduced perceived safety

3 Methods

3.1 Experimental design

To gather experimental data, we conducted a merging experiment wherein participants were instructed to adhere to their typical driving behaviour while engaging in interactions with each other. It is crucial to highlight that participants were informed they would only interact with each other, with no mention of interactions with AVs. Verbal communication was expressly prohibited, and participants were equipped with noise-cancelling headphones to enforce communication restrictions. To get used to the experimental environment, vehicle dynamics, and the methodology for assessing perceived safety, participants underwent a series of ten training trials. Initial conditions were systematically randomized to account for potential learning effects and to mitigate the influence of uncontrolled variables.

The driving experiment featured a first-person perspective within the vehicle, as shown in Figure 1 and encompassed two straight roadway segments, a parallel merging junction, and an additional straight road. Initially, in the straight roadway segment, the vehicles were operating under cruise control, where the participant could only steer the vehicle. At the beginning of the parallel merging junction which occurs at the exact same time and position for both interacting agents, participants assumed control upon visual cues provided by road signage, accompanied by auditory cues from their headphones. The trial ended with the participant being able to fully control the vehicle on the last straight road. Figure 2 shows the complete track from a top-down view.

Figure 1: Participants' Point of View inside of their vehicle during the experiment. In the middle of the steering wheel the participant can see the value of their Perceived Safety

3.2 Setup

The experimental setup involved a virtual reality simulated driving environment. The visual environment of the experiment was developed using Unreal Engine 4.26, CARLA 0.9.13 [24], an open-source simulator specifically designed for autonomous driving research. Additionally, JOAN, an open-source software framework [25], was utilized in conjunction with CARLA.

Participants used a USB steering wheel with pedals to control their vehicle (Logitech Driving Force GT). Additionally, the participants used two buttons on the back of the steering wheel to select the perceived safety value.

The Varjo VR3 VR headset was selected for its high-resolution display, immersive capabilities, and advanced tracking features.

Figure 3 provides a visual representation of the complete experimental setup.

3.2.1 Initial Conditions

The starting positions are characterized by variations in initial lane positions and interaction partners. These positions were designed to allow participants to experience both roles in the merging manoeuvre: as the merging agent ("On-Ramp") and as the agent in the highway lane ("Highway"). Additionally, to compare HV-HV and HV-AV interactions, participants either interacted with another human ("with

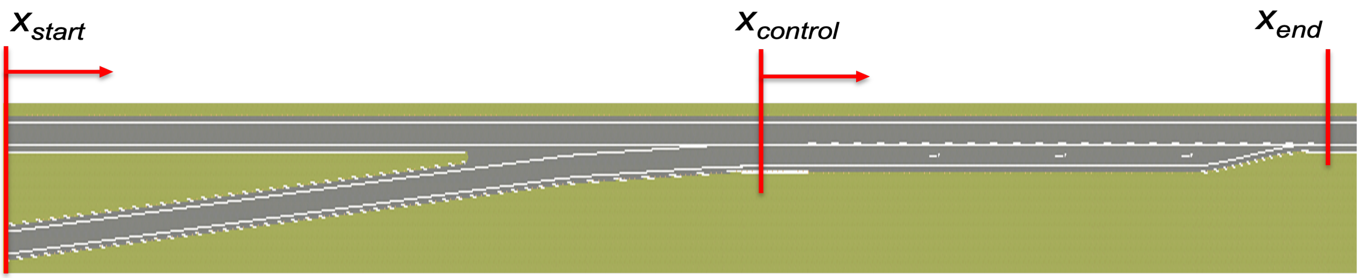

Figure 2: Track used for the experiment, with merging lane and highway lane. Where x_{start} shows the starting positions, $x_{control}$ shows where the participants get control and x_{end} shows the end of the merging lane

Figure 3: Experimental setup. Two participants taking part in the experiment each with a steering wheel, pedals and VR headsets

human" in Table 1) driver or an AV ("with AV" in Table 1). Each participant interacted with both interaction partners in all the lateral positions. For clarity, in our experimental design, the naming of the conditions in Table 1 are primarily focused from the perspective of driver 1. It's important to note that in the condition "Highway with human" the data of driver 2 is concurrently captured and categorized as the condition "On-Ramp with Human".

Each vehicle was designated by a specific colour code (yellow, blue, green, and grey). Yellow and blue corresponded to the vehicles operated by the participants, with participant 1 assigned to the yellow car and participant 2 assigned to the blue car. The AVs' colours were green and grey. Participants interacted with all interaction partners, thus with three different coloured-coded vehicles. After the experimental sessions, participants were asked to complete a questionnaire to describe any behavioural distinctions between the different coloured vehicles.

Each experimental condition was executed ten times per session, organized in a randomized sequence, yielding a total of 40 iterations. Throughout the experiment, the speed limit was consistently set at 80 km/h (kilometres per hour). Participants were explicitly instructed to exceed this speed limit exclusively under critical circumstances.

Condition	Driver 1	Driver 2	AVs	Interaction
Highway with Human	Highway lane	On-Ramp	-	HV-HV
On-Ramp with AV	On-Ramp	On-Ramp	Highway lane	AV-HV
On-Ramp with Human	On-Ramp	Highway lane	-	HV-HV
Highway with AV	Highway lane	Highway lane	On-Ramp	AV-HV

Table 1: Conditions and interaction type from the perspective of Driver 1

3.3 Exclusion Criteria

All 14 (7% of the total data) data gathered from trials with collisions were excluded by the analysis and marked as failed interaction. This is motivated by two main reasons. Firstly, our primary research interest lies in understanding the dynamics of successful AV-HV interactions during merging manoeuvres. Collisions represent extreme cases where the interaction fundamentally breaks down, thereby not providing useful insights into the normal range of interactions we aim to study. Secondly, including collision data could introduce significant outliers and thus affect the results, leading to misinterpretations about typical driver behaviour and interaction patterns. By focusing on non-collision trials, we ensure that our analysis remains consistent and relevant to typical driving scenarios.

Additionally, all the data were analysed only in the time window between the drivers starting to control their vehicle until the merging manoeuvre was successfully completed. This was decided to only analyse the data where the interaction that influences the manoeuvre took place.

3.4 Assessment of human-robot interactions

3.4.1 Merging Outcome

To compare how the high-level outcome of the interaction differs between HV-HV and HV-AV interactions we evaluated the number of collisions and the percentage of merges in front of the other vehicle. The latter metric shows which vehicle was in front when the merging vehicle crossed the dashed middle line.

3.4.2 Workload

Workload is a key parameter used to assess the cost of task completion within human-machine systems [26]. It can be influenced by various factors, including task complexity, time constraints, equipment quality, working conditions, task performance, operator skills, strategies, experience, and perception. There are multiple methods to measure workload empirically.

- **Self-assessment:** Workload is often evaluated through questionnaires, with the NASA-Task Load Index (NASA-TLX) [27] being a common choice. However, it is important to consider potential bias in questionnaire responses [28] due to participants' trust in technology.
- **Secondary tasks:** Another approach involves evaluating workload through performance on secondary tasks, where the ability to perform a secondary task reflects residual workload [29]
- **Psychophysiological methods:** Workload can also be assessed using psychophysiological indicators, including:
  - *Cardiovascular Measures:* Workload can be measured through cardiovascular indicators such as heart rate [30] and heart rate variability [30])
  - *Brain Activity:* Workload can be measured through brain activity (EEG) [31], but this approach may be costly and intrusive
  - *Respiratory Activity:* Workload can be measured through respiratory rate [32]
  - *Eye Tracking:* Eye-tracking technology, readily available in many VR headsets, offers a cost-effective and non-intrusive means of workload assessment. It can provide data on eye movement (e.g., gaze coordinates and fixations)[33–35], blink rate [36], blink duration [37]), pupil diameter [38], making it a recommended option for workload evaluation [33]

Eye-tracking technology provides a non-intrusive and cost-effective way to assess workload. One effective method for collecting eye-tracking data during experimental studies is through the use of VR technology. Previous studies have demonstrated the effectiveness of VR in capturing driver performance data, including eye-tracking metrics [39]. Additionally, the use of VR technology in experiments can significantly enhance participant engagement by creating a more immersive experience. Furthermore, VR offers a cost-effective alternative to more expensive setups, such as moving-base simulators. It is also noteworthy that VR experiments have been found to induce motion sickness in participants, an effect similar to that experienced in traditional simulators [39]. The immersive quality of 3D-VR environments has yielded results comparable to those obtained from other types of simulation environments. The added benefit of integrated eye-tracking capabilities in VR headsets makes 3D VR a more economical choice for conducting experiments.

The driver's gaze angle is an important metric for eye-tracking and it is the sum of the driver's head rotation and eye movement: $\alpha_{gaze} = \alpha_{head} + \alpha_{eye}$. Figure 4 visually represents how the gaze angle is computed.

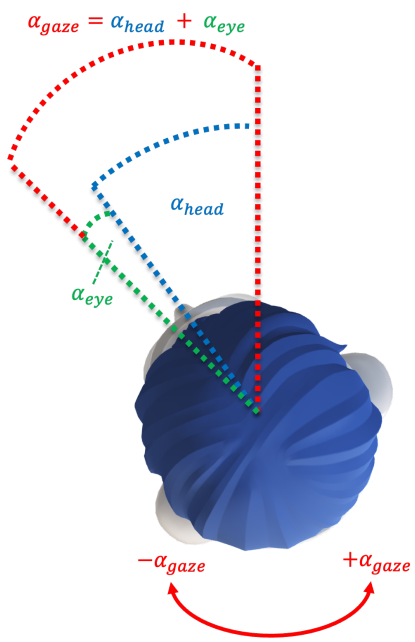

Figure 4: Visual representation of the computation of the driver's gaze. The angle α_{head} represented in blue depicts the head angle relative to the origin. The angle α_{eye} coloured in green represents the angle of the driver's eye, which is relative to the driver's face. Finally in red the angle α_{gaze} shows the gaze angle computed by summing the two aforementioned angles. It is important to note that since the eye angle is relative to the driver's face, with the same head angle we can compute different gaze angles depending on where the driver is looking at. Furthermore, the angles in Varjo and JOAN follow the left-hand rule, the head movement to the left is a negative angle as well as the angle between the cars is negative for the vehicle in the on-ramp

Whenever the absolute value of α_{gaze} exceeded $\alpha_{vehicle1}$ (refer to Figure 5) for Driver 1 ($\alpha_{vehicle2}$ for Driver 2) or exceeded a fixed threshold of 50 degrees (*Threshold for vehicle 1* for Driver 1 and *Threshold for vehicle 2* for Driver 2), it was interpreted as the driver verifying the presence of the other vehicle. It is important to note that any eye-tracking data recorded after the completion of the merging manoeuvre was excluded from the analysis. This exclusion was guided by two considerations: firstly, the primary focus of the study was on the workload related to the merging manoeuvre itself; secondly, fixation patterns on the other vehicle were potentially altered following the completion of a merge by the other car.

The evaluation of workload was based on two key metrics:

1) **Number of times the participants checked the presence of the other vehicle:** A higher number of times is indicative of increased workload [34].
2) **Total time the participants checked the presence of the other vehicle:** Longer duration suggests that drivers require more time to comprehend the behaviour of other cars, signifying a higher workload [34].

3.4.3 Perceived Safety

Safety, a critical aspect of user comfort, is often categorized into physical and perceived (also known as psychological) safety in the literature. In AV-HV interaction where collisions are not accepted as possible outcomes, only perceived safety is considered, which plays a pivotal role in human-robot interaction (HRI) comfort, significantly influencing attitudes and trust in technology.

Psychological safety encompasses perceived safety and stress levels. Typically, questionnaires are employed to assess the user's perceived psychological safety, focusing on aspects such as the absence of obstructions, the ability to maintain preferred velocity in the robot's presence, and overall impressions of the interaction. Common questionnaires include the Godspeed questionnaire [40], the Robotic Social Attributes Scale (RoSAS) [41] (derived from the Godspeed questionnaire), and the BEHAVE-II instrument [42].

Combining questionnaires with real-time tracking of physiological signals like heart rate, heart rate variability, and eye gaze is often recommended [8]. Heart rate is considered a crucial biomarker linked to the autonomic nervous system's activation, providing insights into stress and fear. Eye gaze monitoring is valuable since individuals tend to focus less on potential sources of danger when they feel safe. It can also be combined with pupillary dilation measurements [8].

Beyond questionnaires and physiological signal evaluation, alternative methods for assessing perceived safety include the use of direct input devices [8]. Direct input devices, resembling joysticks, serve a purpose similar to questionnaires in terms of data collection during experiments. However, they offer a distinct advantage by providing real-time feedback, which is not possible with questionnaires. Nevertheless, it must be noted that these devices can potentially be distracting and invasive for the driver. Therefore, their design and suitability for the driving experiment are paramount considerations.

In this work, perceived safety will be analysed through direct input devices. During the experiment, participants continuously self-evaluated this metric using the gear pads on the steering wheel. The assessment method was based on a "tick event" approach, where any event altering their safety perception prompted an update of the perceived safety value. This value was also visually displayed on the steering wheel, typically in the location of the horn, ensuring that drivers had constant visibility of their safety status.

To compare how the participant perceived the situation, the drivers' perceived safety at the exact moment when the merging vehicle crossed the dashed line was compared. This focus was chosen for several reasons. First, the conclusion of the merging manoeuvre represents an important, if not the most important point of the interaction, reflecting the outcomes of the merging manoeuvre. Second, by standardizing the analysis to this specific moment, we can more accurately compare perceived safety levels across different trials by uniforming the metric. The evaluation scale employed for this metric was -1, 0, and 1, categorizing perceived safety into three states: unsafe (-1), normal (0), and safe (1).

3.4.4 Questionnaire

After the trials, each participant was asked to fill in a questionnaire. This was done with the rationale of assessing if the drivers noticed any difference in behaviour between the colour-coded vehicles. The following questions were asked:

1) To which extent were you able to understand the other participant's intentions?
2) Have you noticed any differences in behaviour between cars of different colours?
3) How large were which differences in behavior that you noticed?
4) How exactly did the behaviour of the other cars differ depending on their colour?
5) How did the different behaviours of the other cars influence your interaction with them?

Question 1 was answered using a scale of 1 to 5 with 1 being "I could not understand it at all" and 5 being "I was able to understand very well". Furthermore, only the participants who answered "Yes" to question 2 were able to answer the remaining questions. Again, question 3 was to be answered based on a scale from 1 to 5, with 1 standing for "Very minor differences" and 5 standing for "Major differences", while the last two questions were open-ended.

Although participants were told they would interact only with each other subsection 3.1, we asked through the questionnaires if they noticed any difference in driving behaviour. Although this might be unintuitive, the expectation was that if participants could independently notice differences in driving behaviours without prior knowledge of the type of vehicle they were interacting with, it would suggest that there are observable and distinguishable behavioural patterns associated with AVs compared to HVs. For this reason, the use of colour-coded vehicles was a deliberate methodological choice, serving a primary purpose; in instances where a vehicle exhibited particularly notable behaviour, the colour association was intended to aid participants in recalling and linking these behaviours to specific vehicles during the post-experiment questionnaires. If the driving behaviour of the interacting vehicles would not differ extremely, the colour association was probably disregarded. Additionally, not mentioning the participants the AVs served different purposes. Firstly, the experiment aimed to simulate a real-world scenario where drivers might

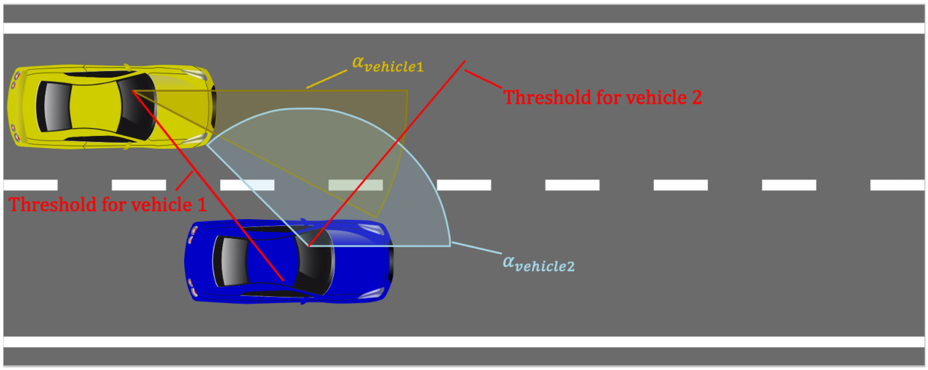

Figure 5: The angle of the two vehicles is computed by taking the driver's position inside of his/her vehicle to the first part of the other vehicle where the driver can see if the vehicles are parallel. For the car on the highway lane (yellow car in this example) the angle $\alpha_{vehicle1}$ is between the driver's position and the front left part of the blue vehicle. Conversely, for the vehicle in the on-ramp, the angle $\alpha_{vehicle2}$ is computed between the driver's position and the most front-right part of the other vehicle. In addition to the angles $\alpha_{vehicle1}$ and $\alpha_{vehicle2}$ a threshold for the driver's gaze is introduced. Both thresholds, *Threshold for vehicle 1* and *Threshold for vehicle 2*, have a value of 50°.

not always know if they are interacting with an AV or an HV. Secondly, the setup allowed for the observation of natural responses to behavioural cues rather than responses influenced by participants' attitudes or beliefs about AVs. Thirdly, the experiment aimed to prevent the participants from consciously or subconsciously seeking out differences between AVs and HVs. The concern was that if participants were informed about the involvement of AVs, they might actively look for distinctions, which could lead to biased observations or exaggerated perceptions of differences that, under normal traffic conditions, might not be noticeable or significant.

3.5 Controller's Adjustments

For this specific experiment, we modified the model of the car's dynamics in [1] to suit Unreal Engine, the simulator utilized in this study. Precisely, we replaced the point-mass model with a bicycle model to allow integration in Unreal Engine. Furthermore, through testing in Unreal Engine, we computed the maximum acceleration and deceleration values alongside the maximum steering angle. We integrated these constraints into the model of the car dynamics primarily for algorithm validation. This validation ensures the accuracy and effectiveness of the AV's controller by confirming its ability to respond appropriately to the vehicle's capabilities within the simulated environment. Furthermore, these constraints play a role in enhancing realism within the simulation by adding authenticity to the simulation and thereby providing a reliable testing ground for the AV.

Once the model was adjusted to the simulator, the values of the reward function weights were estimated. This is because the whole environment for the algorithm simulation changed from a 2D world into a first-person view 3D world (Figure 1) with different physics. It is important to note that both adjustments were estimations and estimated through several trials. Following the implementation of the algorithm, we designed the experimental track and environment around the integration and evaluation of the algorithm's interaction performance. Finally, we implemented the controller which was executed on a Linux computer connected to the simulator server.

3.6 Statistical Analysis

Statistical significance for Workload and Perceived Safety was assessed using Mixed Effects Module between the conditions. We set the significance level for the statistical analysis at $p < 0,05$. All statistical analyses were conducted using Python.

3.7 Participants

The experiment was structured into ten sessions, each involving a randomly assigned pair of participants, resulting in a total of 20 participants (10 pairs). The selection criteria for participants included the ability to drive without the use of corrective glasses or contact lenses, as this was necessary for the use of Virtual Reality Headsets, and the possession of a valid driver's license. All the participants gave informed consent to do the experiment and were compensated with a €15 gift card for their participation. The average age of the participants was 25,7 years with a standard deviation of 4,23 years. Six participants (30%) identified themselves as female while 14 (70%) identified themselves as male. The average years of possessing a driving license was 7 years with a standard deviation of 3,78 years. Most of the participants were students and researchers of TU Delft, specifically of the Faculty of Mechanical Engineering and thus already familiar with the concept of autonomous vehicles, autonomous vehicles' algorithms and the field of Human-Robot Interaction. The experiment received ethical approval from the ethical committee of TU Delft.

3.8 Data

To assess all the aforementioned metrics, many different data were recorded through JOAN and the Varjo headset.

3.8.1 Vehicle Data

We collected the vehicles' data through JOAN. We collected the timestamps, the drivers' inputs through the steering wheel such as brake, gear, steering angle and throttle, as well as the vehicle data such as acceleration and velocity (both in Vehicle frame and in World frame), and the vehicles' transform comprehending position (X, Y, Z) and angles

(Yaw, Pitch, Roll). Joan collects data with a frequency of 100HZ.

Due to technical reasons, the value for Perceived Safety changed by the participants was saved under "gear input" in JOAN.

3.8.2 Eye Tracking Data

We used the native Varjo SDK for all the eye-tracking data. We measured the timestamps, the participants' head movement and the participants' eye coordinates. Through the native SDK, data is gathered at a frequency of 100HZ.

4 Results

We compared the condition in which participants merge in a human-human (HV-HV) interaction to the condition in which they merge in a human-autonomous vehicle (HV-AV) interaction. Additionally, we compared the condition where participants were in the highway lane in an HV-HV interaction to the condition where they were in the highway lane in an HV-AV interaction. In summary, we were comparing condition "Highway with human" to "Highway with AV" and condition "On-Ramp with human" to "On-Ramp with AV". This comparison allowed us to evaluate the workload and perceived safety associated with the same manoeuvres (merging or staying in the lane) in both HV-HV and HV-AV interactions.

4.1 Merging Outcome

4.1.1 Number of Collisions

The total number of collisions was 14 which represents the 1,75% of the total trials.

In scenarios involving human drivers, whether in the "Highway" or "On-Ramp" conditions, collisions were the least frequent, occurring in only 2 and 1 instances out of 200 trials (refer to Table 2), respectively. This suggests a relatively low impact on overall safety in human-to-human interactions.

On the other hand, interactions with the AV resulted in a higher number of collisions, with 5 and 6 collisions out of 200 trials for "Highway" and "On-Ramp," respectively.

	Highway with human	*Highway with AV*	*On-Ramp with human*	*On-Ramp with AV*
Total trials	200	200	200	200
Total collisions	2	5	1	6
Collision rate	1%	3%	1%	3%

Table 2: Total number of collisions and collision rate per condition

There is no statistically significant difference between the number of collisions in the different conditions ($p = 0,494$ and $p = 0,055$, respectively).

4.1.2 Percentage of Merges in front of the other Vehicle

There are differences in the results of the percentage of merges in front of the other vehicle between the conditions "Highway" and "On-Ramp". In the "Highway" conditions, with both interaction partners "with human" and "with AV", the vehicle already on the highway stayed in front most of the time and let the merging vehicle merge behind. The difference between these two conditions is relatively small, in the "Highway with human" condition, 79% of the time the vehicle in the highway was in front when the merging vehicle merged in, while in the condition "Highway with AV", this happened 72% of the time (refer to Figure 6). These results led to no statistically significant difference ($p = 0,422$).

Obviously, in "On-Ramp" scenarios with both interaction partners, "with human" and "with AV", the merging vehicle tends to merge in front of the other vehicle fewer times, 21% and 34%, respectively (refer to Figure 6). In contrast to the "Highway" conditions, the difference between these conditions is statistically significant ($p = 0,024$).

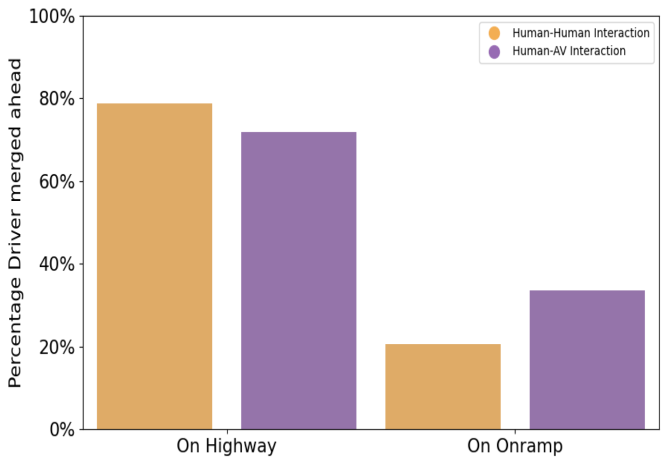

Figure 6: Percentage of merges in front of the other vehicle per condition

4.2 Workload

Figure 7 shows a visual representation of how the data was analysed for each repetition in each trial of the experiment. The plot shows the data gathered for Driver 1 in the condition "Highway with Human". This visual representation aids in comprehending how and when drivers were monitoring the presence of the other vehicle during the trials. In this example, we can see that the driver checked the other car's presence during the cruise control twice and, more importantly, several times (4) after taking control of the vehicle. Furthermore, it indicates a decrease in the angle between the vehicles, suggesting a change in their relative positions due to speed variation. This leads to the other vehicle merging in front, as shown by an approximately 5-degree angle at the merging point.

In our results, participants checked more times the other vehicle's presence when interacting with the AV in the "Highway" condition, meaning that the way the AV merges and interacts while merging necessitates more glances (refer to Figure 8). For the condition "On-Ramp" there doesn't

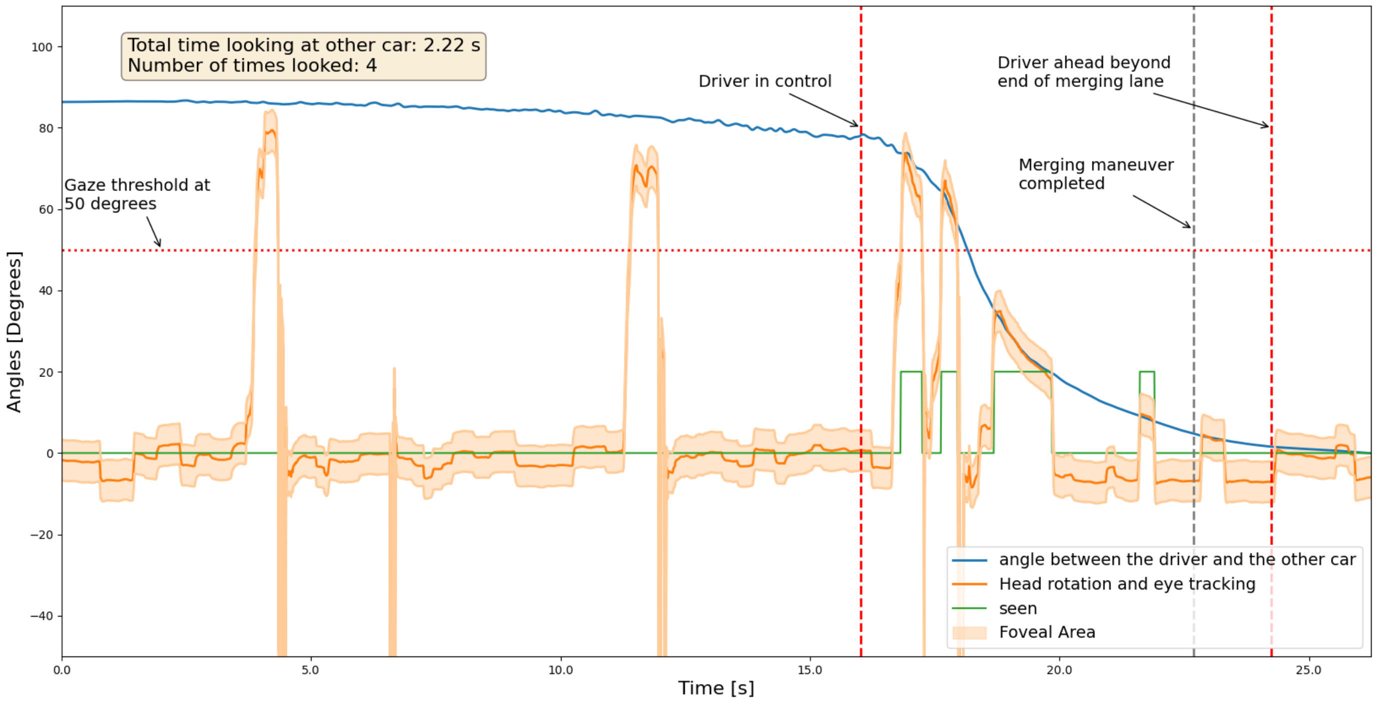

Figure 7: The row data of the driver's eye tracking. The orange line shows the driver's gaze angle and her/his foveal region, and the blue line stands for the angle between the two vehicles, in this case, the angle is positive because the condition was "Highway with Human" and thus the recorded data shows the eye-tracking for the driver on the highway. The green line shows if the driver checked the other's vehicle presence, every time the driver's gaze angle was higher than the thresholds or higher than $\alpha_{vehicle1}$ the green line has a value of 20. This line is only used to have visual feedback on the data. As we can see from the plot the driver was checking the other vehicle's presence for about 2,22s, sometimes by looking at the other vehicle directly and sometimes only by checking its presence.

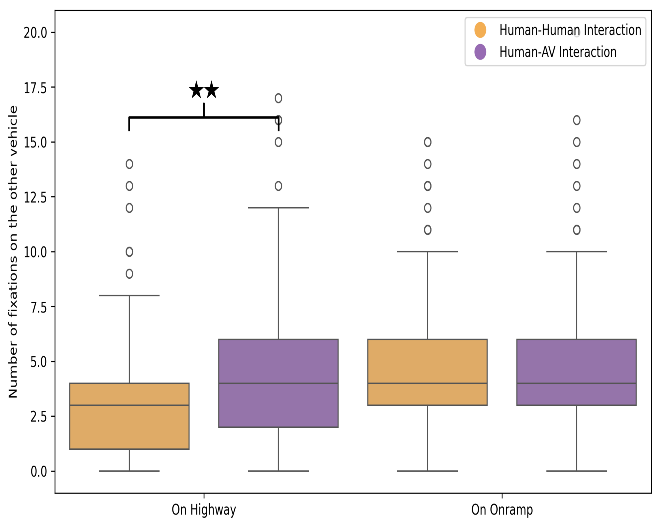

Figure 8: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type

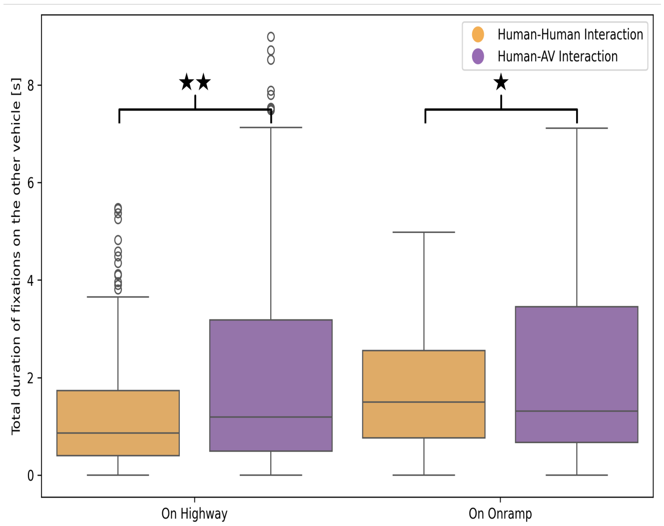

Figure 9: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type

appear to be any difference in the number of times the participant checks the other vehicle's presence (refer to Figure 8). This might suggest that the AV is better at interacting when it is in the highway lane. We saw a difference in the total time spent by the driver looking at the other vehicle in both conditions "Highway" and "On-ramp" when interacting with the AV compared to interacting with the HV (refer to Figure 9). These results might suggest that the driver needs more time to understand the AV in both conditions.

The statistical analysis did reveal a significant difference ($p = 0,002$) only between the conditions "Highway with human" and "Highway with AV" in terms of workload, which was measured based on the frequency of drivers checking the presence of other vehicles (Table 3). As expected, when looking at Figure 8 the frequency of drivers checking the presence of other vehicles is significantly higher with the AV merging compared to when it was another HV that was

Conditions	z	P-value> | z |
"Highway with human" and "Highway with AV"	3,056	0,002
"On-ramp with human" and "On-ramp with AV"	0,948	0,343

Table 3: Mixed Effects Module test for Workload measured through the number of times the driver was checking for the other's vehicle presence between conditions "Highway with human" and "Highway with AV" and between conditions "On-Ramp with human" and "On-Ramp with AV"

merging. This means that the drivers had to glance more often towards the AV when it was merging. There is no statistically significant difference in the frequency of drivers checking the presence of other vehicles when comparing the two "On-Ramp" conditions.

Conditions	z	P-value> | z |
"Highway with human" and "Highway with AV"	2,429	0,015
"On-Ramp with human" and "On-Ramp with AV"	2,012	0,044

Table 4: Mixed Effects Module test for Workload measured through total time spent by the driver on checking for the other's vehicle presence between conditions "Highway with human" and "Highway with AV" and between conditions "On-Ramp with human" and "On-Ramp with AV"

The analysis of workload measured through the total time spent by the driver on checking for the other's vehicle presence indicates a statistically significant difference between conditions "Highway with human" and "Highway with AV" ($p = 0,015$) and between conditions "On-ramp with human" and "On-Ramp with AV" ($p = 0,044$) (Table 4).

The results for workload confirm the hypothesis that drivers need more time to understand the AV's behaviour in a merging scenario.

4.3 Perceived Safety

Perceived Safety data was captured as a continuous signal throughout the experimental trials. Figure 10 illustrates how this raw data was collected during a representative trial, specifically in the 'On-Ramp with AV' condition. The figure illustrates a shift in the driver's perceived safety, highlighting a marked decrease once the merging process started. This visual representation underscores the interaction's impact on the driver's perceived safety, particularly at the moment when the merging vehicle crosses the middle dashed line.

The results of our study showed that drivers generally felt less safe when interacting with the AV (refer to Figure 11). The average Perceived Safety was lower when participants were interacting with the AV, especially in the condition "On-Ramp", suggesting a perception of lower safety in this condition.

Figure 12 presents bar plots illustrating the proportions of time participants felt safe, neutral, or unsafe at the aforementioned crucial moment across all conditions. The plot distinctly reveals that participants felt unsafe more frequently in "with AV" conditions than in "with human" conditions at the critical moment of merging manoeuvre

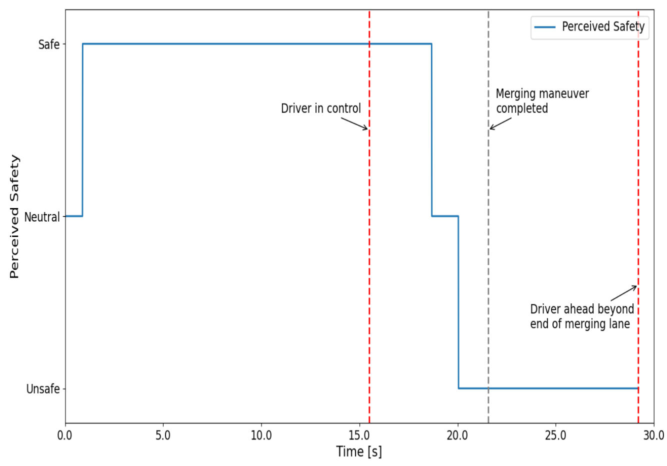

Figure 10: Raw data of Perceived Safety for a representative trial ('On-Ramp with AV' condition). The continuous blue line represents the Perceived Safety signal across the entire trial. The first dashed red line indicates when the driver assumed control of the vehicle, and the second dashed red line marks the point where the driver crossed the end of the merging lane. The dashed grey line signifies the completion of the merging manoeuvre

completion (red bars). Additionally, the drivers felt "safe" fewer times during interaction with AV compared to interaction with HV (green bars), suggesting AV interaction or behaviour is perceived more negatively than that of HVs.

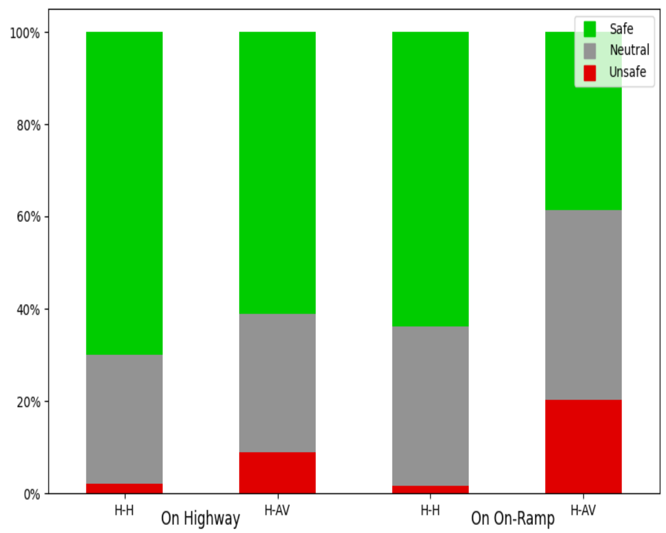

Figure 12: Proportion of perceived safety when the merging vehicle crossed the middle line.

To compare the results, we summarized the aggregated instances where participants reported feeling unsafe at the aforementioned moment across each experimental condition. This approach was crucial in capturing the immediate perceived safety resulting from the merging manoeuvre.

The statistical results (Table 5), indicate a statistically significant difference in the "On-Ramp" conditions ($p = 0,002$), whereas no significant difference was found between the "Highway" conditions ($p = 0,099$).

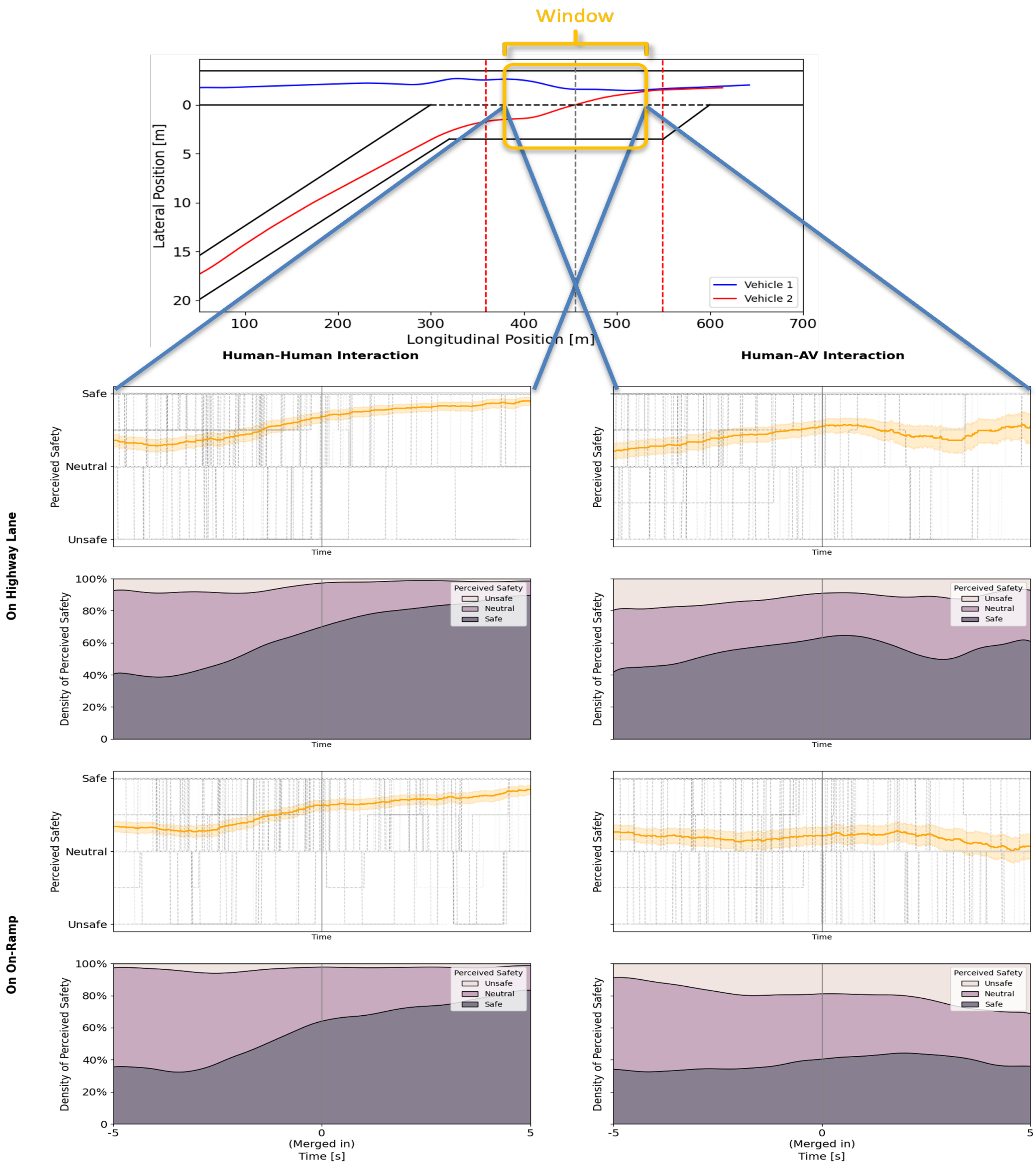

Figure 11: Perceived Safety and its density for all 10 sessions, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window (in yellow), to have a better understanding and comparison of the interaction at the same time. The plot above shows the lateral and the longitudinal positions of the vehicles during one representative trial ("Highway with human"), it also shows the fixed time window in the global picture of one representative trial.

Conditions	*z*	*P-value>∣z∣*
"Highway with human" and "Highway with AV"	1,651	0,099
"On-Ramp with human" and "On-Ramp with AV"	3,036	0,002

Table 5: Mixed Effects Module test for Perceived Safety between conditions "Highway with human" and "Highway with AV" and between conditions "On-Ramp with human" and "On-Ramp with AV"

4.4 Questionnaires

The questionnaires aimed to measure if drivers perceived behavioural differences across colour-coded vehicles and connect these observations with objective data, enriching the study's insights with personal feedback.

Participants generally rated their understanding of other drivers' intentions as neutral, with an average score of 3.5 and a standard deviation of 0.69. A majority, 65% of participants (13 out of 20), observed behavioural variances between different coloured cars, with 61.6% considering these differences substantial (average rating of 3.75 and a standard deviation of 1.22).

To better understand the following answers is good to remember that the green and grey vehicles were AVs, while the blue and the yellow ones were HVs. The The question *"How exactly did the behaviour of the other cars differ depending on its colour?"*, three main descriptors noticed:

1) ***Not legible***: Participants had difficulty in understanding the AV's behaviour. Its implication is well summarized by one participant's answer: *"the green carfor example didnt know what he wanted to do, accelerate or stop so that was annoying"* (participant 2). Another participant stated their difficulty in understanding the other vehicle's behaviour without specifying a colour-coded car *"abrupt speed changes, non predictable driving behaviour"* (participant 15).
2) ***Technical driving***: Some participants complained about the technical driving capability of the AVs, especially on how smooth the other vehicle felt, by stating *"grey car not seem to accelerate smoothly, but that might also have been the other person driving"*(participant 12), or *"When merging, the green and grey cars would slow down, skid"* (participant 4). Others didn't specify any colour but only stated noticed differences, so the following statements are not clearly directed towards AVs: "*abrupt speed changes*" (participant 15) and "*sometimes abrubtly braking*" (participant 8).
3) ***Driving behaviour***: Different participants described the other vehicle's behaviour without explicitly stating the colour *"Some colours seemed to behave more reckless and aggressive, sometimes abrubtly braking"* (participant 8), *"The driving style (more or less aggressive)"* (participant 16) and *"I don't know how they were different but I saw different behaviours and expect them to be related to the colours"* (participant 18). One participant (participant 20) clearly stated the difference in behaviour by mentioning also the colour of the vehicles *"Yellow - normal behaviour; Green or grey - aggressive, reckless, trying to get ahead"*.

In general, we can identify at least one of these three descriptors in ten out of 13 participants. These different results yield that, in our study, the majority of people see differences in AV driving behaviour, though there is no general consensus on what was really differing. Participants expressed frustration with the AVs' unclear intentions and abrupt speed changes, indicating a need for improvement in AV communication and behaviour predictability to enhance safety and acceptance among human drivers.

The answer to the final question *How did the different behaviours of the other cars influence your interaction with them?* can be divided into two main groups:

1) ***Proactive Participants***: Those who took the initiative by speeding up: *"Speed up and merge in front of them"* (participant 4), or overtaking, especially when foreseeing potentially dangerous behaviour from other vehicles: *"when I could foresee a dangerous behaviour because of the colour of the car I would tend to take initiative to overtake it so that its behaviour and speed changes does not affect me"* (participant 15)
2) ***Cautious Participants***: Those who responded more conservatively, maintaining larger distances: *"Larger distance = safer perception"* (participant 5) or waiting for the other vehicle to act first: *"waited for the action of the other participant"* (participant 7), especially when perceiving the behaviour as reckless or aggressive: *"was more cautious when the behaviour seemed more reckless"* (participant 8). Participant 20 summarises: *"Was much more careful around the green and grey cars, let them go ahead"*

Five participants out of 13 reported that they took the initiative, while 7 participants described their reaction as more conservative. None of the participants reported a complete lack of change in their behaviour in response to the change in the behaviour of other vehicles. However, one participant did mention that their behaviour was only slightly influenced, stating: "*Not that much, I was more careful around the grey car*". Interestingly, the participants were nearly evenly split between these two reactive approaches, five participants out of 13 took the initiative while seven out of 13 acted more cautiously, and one participant didn't express a clear answer. This spread highlights the diverse strategies in response to perceived differences in the behaviours of colour-coded vehicles. The results show that more than half of our participants saw differences in behaviour, with many describing the AV driving behaviour negatively.

5 Discussion

We conducted a driving experiment to assess the difference in interaction between the HV-AV interaction and the HV-HV interaction through high-level outcomes, workload and perceived safety, and to check if human drivers can recognise differences in driving behaviour between autonomous vehicles and human-driven vehicles. Workload was measured through two different metrics: the number of times the participants checked the presence of the other vehicle and the total time the participants checked the presence of the other vehicle. We discovered significant differences both in workload and in perceived safety in both conditions for workload and in the "On-Ramp" condition for perceived safety, respectively. Furthermore, the answers to the questionnaires reveal that human drivers perceive the evaluated AV driving behaviour as different from an HV. Additionally, the outcomes of the questionnaires provide valuable insights into the participants' perceptions and reactions during the experiment and give a deeper understanding of the two previously mentioned metrics.

The findings indicate a notable distinction in driver workload and perceived safety when interacting with AVs compared to when they were interacting with HVs. The increase in workload, expressed as the total duration of drivers checking the presence of the other vehicle, suggests that AVs may require drivers to allocate more attention to understand their behaviour, potentially leading to increased cognitive demands [34]. Similarly, the decreased perceived safety in "On-Ramp with AV" conditions highlights possible concerns about the predictability and smoothness of AV driving patterns, affecting driver perceived safety and drivers' reactions to their behaviour as seen from the results of the questionnaires. The findings cannot be generalized to all AV controllers, but, together with the lack of AVs' interaction evaluation, highlight the necessity of a new and different AV evaluation to identify and mitigate any negative impacts of AVs on human drivers. Additionally, Figure 11 shows a difference in drivers' perceived safety across the entire trial between these conditions and not only at the end of the merging manoeuvre, meaning that the whole interaction with AVs is badly influenced by AV-driving behaviour.

The concerns raised by participants about the predictability and comprehensibility of the AVs could be a contributing factor to the increased workload observed during interactions with AVs. As drivers expend more effort to understand the intentions and behaviours of AVs, there is a rise in workload metrics we've previously covered. Additionally, the perceived behaviours of AVs as *"non predictable"*, *"more aggressive"* and *"reckless"*, as described by the participants, could account for the reduced perceived safety in the "On-Ramp with AV" condition. This might suggest a direct correlation between the participants' subjective experiences and the quantitative findings of the study, highlighting the impact of AV behaviour on human driver responses.

These results have profound implications for the design and operation of AVs. They underscore the necessity of incorporating human factors, such as workload and perceived safety, into AV controllers design to ensure that these systems do not adversely affect the driving experience.

The results highlight a critical aspect of AV controllers' current design, namely, the need to ensure that these systems are assessed also with a comprehensive evaluation of their impact on their interactions with HVs. This underlines the critical importance of incorporating safety perception in AV algorithm design, as the success of AV integration relies heavily on driver trust and acceptance. It is crucial for AVs to be perceived as safe by human drivers to facilitate their integration into current road systems and, even more so, in future mixed traffic scenarios. If AVs are perceived as unsafe, their acceptance by human drivers could be compromised, limiting their benefits to AV users only. This situation could hinder the widespread adoption and utility of AVs in existing traffic ecosystems. Therefore, ensuring that AVs are perceived as safe by all road users is essential for their successful integration and functionality in modern transportation. The findings underscore the complexity of human-AV interactions and highlight the need for designing AVs that are intuitive and predictable to human drivers. Additionally, the diversity in human responses also emphasizes the importance of considering various human behavioural patterns when developing and testing AV algorithms, leading again to the importance of the human factor when designing AVs. Lastly, our findings consistently show that in merging scenarios, participants were always affected by the behaviour of the AV underscoring once again, the impact of AV driving behaviour on human drivers. This, again, highlights the need for careful assessment of AV-HV interaction during the design phase of AVs. Ensuring that AVs are designed with a keen understanding of their impact on human drivers is essential for their successful integration into the transportation ecosystem.

Having highlighted the implications of our results, we also reflected on the use of eye tracking to evaluate workload. Under workload, one can find different definitions - cognitive, mental and physiological - in more recent studies, all fall under the definition of driver's workload (DW) [43]. Researchers used different methods to measure DW. Eye tracking, in particular, has been utilized in different ways to assess workload, including blink rate [36], blink duration [37]), pupil diameter [38], and fixations [33–35], all recognized as significant indicators of DW [44].

Workload is already studied on AV and depending on its level, automation has been shown to impact driver workload negatively [45, 46] and positively [33].

Although used for workload, eye tracking, in the field of driving, is also employed to measure situation awareness (SA) [27, 47]. In this case, the focus is not just on the cognitive load but on whether and how well drivers perceive, comprehend, and anticipate environmental elements [48]. Unlike the specific application in measuring workload, situation awareness assessment encompasses a broader understanding of the situation. In other words, applied to our experiment, SA would indicate if the drivers understand that: a) the other vehicle wants to merge in, b) how it will merge in and c) what will be the future manoeuvres. However, DW, in our experiment, determines how much time the driver needs to understand the aforementioned SA.

Most existing research on eye tracking and workload, especially concerning automation, does not directly address the impact on human drivers interacting with AVs. This study innovates by focusing on the effect of AV automation on the workload of HV drivers through eye tracking.

Given the simplified nature of our experimental design, featuring only two vehicles on the track without the presence of non-driving related tasks (NDRT) that might divert the driver's attention, our focus shifts to quantifying the time required by drivers to comprehend the intentions of the other vehicle and to respond accordingly to manoeuvre successfully. This approach aligns with our objective of measuring workload, emphasizing the cognitive effort required from drivers, as measured by the time they need to invest, rather than assessing their overall situation awareness. Furthermore, the consistent conditions across trials should imply that any differences detected via eye-tracking metrics, particularly fixations, reflect the workload imposed on drivers rather than their understanding of the situation. Hence, in this context, eye tracking serves as a tool to quantify DW between HVs and AVs, marking a novel approach in the exploration of automated vehicles' impact on human drivers.

Finally, the data gathered during the experiment demonstrated that the AV was capable of executing merging manoeuvres and allowing other vehicles to merge, without resulting in a statistically significant higher number of collisions. Additionally, the high-level dynamic of the merging manoeuvre in terms of the percentage of merges in front of the other vehicle showed only a slightly significant difference in the "On-Ramp" conditions. If we had assessed the AV only on these two metrics, our findings would suggest that the AV can execute merging manoeuvres with an "apparently" comparable level of safety to human drivers. If only relying on these metrics, one might infer a comparable safety level between AVs and human drivers. However, a comprehensive examination of additional metrics uncovers substantial differences (as discussed above) that negatively impact AV-human driver interactions.

6 Limitations and Future Works

Our study faced several limitations, including the absence of rear-view mirrors in the VR setup, forcing participants to rely on head movements for environmental assessment. Due to the complexity of the setup (see Figure 3), the addition of reflective surfaces such as rear-view mirrors could lead to network lag and delayed visuals in the VR headsets resulting in motion sickness for participants, thus making the experiment unfeasible. Feedback from participants suggested that the lack of mirrors deviated from real-world driving conditions and that the physical burden of wearing a VR headset was tiring which may have caused a reduction in head movements, especially towards the latter part of the experiment, influencing our results. Since the conditions were randomized there might be cases where some conditions were repeated mostly towards the end of the session. Thus, perhaps, participants were influenced greatly by the physical burden of wearing the VR headset.

Our study also raises questions about the effectiveness of using eye-tracking as the only method to evaluate drivers' workload in quick manoeuvres such as merging. Considering the brevity of these manoeuvres, eye-tracking may not capture all the differences in workload effectively. Alternative methods, as discussed in [49], such as brain activity measured through electroencephalogram (EEG), could provide deeper insights. EEG has been proven effective in various scientific experiments for assessing workload. Nonetheless, its application alongside VR technology may pose challenges, given the invasiveness and physical discomfort associated with wearing multiple head-mounted devices. This concern was raised in feedback from participants about the VR headset's weight in our study. Future work could also use other methods such as measuring cardiovascular activity, skin conductance, and respiratory rate. These methods, compared to EGG may offer a balance between accuracy and participant comfort, especially in immersive VR environments. By assessing workload through cardiovascular activity, skin conductance, or respiratory rate, the experiment could take place in a simulator with a moving base which is comparable with real traffic scenarios [50], to have an even better reality-like experience of HV-AV interactions.

An additional limitation of this work is that we analysed only one controller from the wide array of controllers discussed in the literature, preventing us from generalizing our results to all controllers. However, we note that the literature reviewed [1, 11–19] lacks evaluation of controllers based on their interaction with HVs, a gap our study begins to address. By highlighting this research gap, our study not only begins to bridge the gap in understanding AV-HV interactions but also underscores the importance for future research to prioritize the evaluation of AV controllers in the context of their interaction with human drivers, ensuring safer and, for the driver, less cognitive demanding integration of autonomous vehicles. Moreover, the experiment faced challenges in achieving perfect synchronization within the simulator. This could be attributed to the fact that participants were interacting with two computers connected over the network. Additionally, the Linux computer responsible for controlling the AV vehicles was also connected over the network, potentially leading to minor delays in the simulation. This occasionally resulted in slightly misaligned lateral positions of vehicles at the moment of driver takeover in some conditions. Such synchronization issues may have influenced the precision and repeatability of our analysis regarding vehicle interactions and drivers' responses.

The trade-off encountered in the questionnaire design, which involved the use of colour-coded vehicles without disclosing the presence of AVs to participants, introduces a limitation that could affect questionnaire outcomes. Specifically, this approach might have created a "human-like" driving behaviour expectation which led to an accentuation of any "non-normal" behaviours, making them more noticeable to participants. This heightened awareness could lead participants to inadvertently emphasize anomalies, influencing the overall perception and evaluation of driving behaviours observed during the experiment. To address this limitation and further understand the dynamics of human-AV interaction, the existing experimental setup could be augmented through a comparative study. This would involve explicitly informing a subset of participants about the presence of AVs while keeping others unaware. This method allows for a direct assessment of how prior knowledge about AVs influences participants' perceptions and behaviours. By analyzing the differences between the responses of informed and uninformed participants, we can also determine the impact of awareness on the detection and interpretation of driving behaviours.

This study also had a limited participant pool and most of the participants were of TU Delft, specifically of the Faculty of Mechanical Engineering and thus already familiar with the concept of autonomous vehicles, their algorithms and the field of Human Robot Interaction. Future research should include a more diverse pool of participants.

In future research, a broader approach that encompasses all moments of perceived "unsafety" could provide more insightful results, since exploring the overall signal profile during the entire interaction may offer a more complete understanding of drivers' perceived safety. Another approach might be to consider all the situations that are "not safe" rather than focusing only on "unsafe" situations. This might already yield interesting results. It is also worth considering other methodologies such as heart rate [8], and think-aloud methods for gathering this metric. These alternative approaches are recommended for exploration in subsequent studies to deepen our understanding of perceived safety in AV interactions.

Future research should focus not only on exploring diverse methods for assessing workload and perceived safety, but also on considering different traffic scenarios, and examining how demographic factors, such as participants' age, gender and experience influence human-AV interaction.

Finally, this research has concentrated on two parameters - workload and perceived safety - demonstrating their critical role in the assessment of AV interactions with human-driven vehicles. The findings underscore the need for AV algorithms that take into account these human factors.

Future studies should extend this research by exploring the additional evaluation parameters in [49] and assessing various AV controllers. Such comprehensive evaluations are essential to ensure the integration and acceptance of AVs in our existing road systems, ultimately enhancing the safety and efficiency of our transportation ecosystem.

7 Conclusion

In this study, we analysed the interaction dynamics between human-driven vehicles and autonomous vehicles in merging scenarios, by evaluating the high-level outcome of the merging manoeuvre, drivers' workload and perceived safety. Based on the experimental conditions and the analyses that integrate both quantitative and qualitative data, we assessed and concluded the following:

1) **Merging Outcomes**: Overall, collisions occurred in 1,75% of total trials, with fewer incidents in HV-HV interactions compared to HV-AV interactions. While no significant differences were found between collision rates in different conditions, merging behaviours varied, only in "On-Ramp" scenarios
2) **Higher Workload in AV Interactions**: The investigation demonstrates an increase in driver workload when interacting with the AV presented in [1] in a merging scenario, characterized by a higher duration of checks for the presence of other vehicles.
3) **Lower Perceived Safety in AV Interactions**: Data indicate a significant reduction in perceived safety during interactions with the AV presented in [1] in a merging scenario, especially when the human driver had to merge in, underscoring concerns over the predictability and legibility of AV behaviour from a human driver's perspective.
4) **Perception of AV Driving Differences**: The questionnaire results reveal that although participants were told that they would interact with each other, they perceived the AV's driving behaviour as different. Furthermore, the different driving behaviour of the AV, influenced the human driver since the AV presented in [1] was considered unpredictable compared to human drivers.

These findings highlight the important need for integrating and thus evaluating human factors considerations, such as workload and perceived safety, into the design of AVs to ensure their successful integration into our current traffic ecosystems. Additionally, these conclusions offer important insights for the ongoing development and evaluation of AV technologies, highlighting the importance of human users' perceptions during their assessment phase.

8 Acknowledgement

The author would like to thank Olger Siebinga and Arkady Zgonnikov for their continuous support and insightful inspiration. The author would also like to thank Samir Mohammad (Delft University of Technology) and Riccardo Sepe (Polytechnic University of Turin) for pre-piloting during the experimental design phase.

References

[1] Dorsa Sadigh et al. "Planning for cars that coordinate with people: leveraging effects on human actions for planning and active information gathering over human internal state". In: *Autonomous Robots* 42 (2018), pp. 1405–1426.

[2] Peter A Hancock, Illah Nourbakhsh, and Jack Stewart. "On the future of transportation in an era of automated and autonomous vehicles". In: *Proceedings of the National Academy of Sciences* 116.16 (2019), pp. 7684–7691.

[3] Dimitris Milakis, Bart van Arem, and Bert van Wee. "Policy and society related implications of automated driving: A review of literature and directions for future research". In: *Journal of Intelligent Transportation Systems* 21.4 (2017), pp. 324–348. DOI: 10.1080/15472450.2017.1291351. eprint: https://doi.org/10.1080/15472450.2017.1291351. URL: https://doi.org/10.1080/15472450.2017.1291351.

[4] Arnab Bose and Petros Ioannou. "Evaluation of mixed semi-automated/manual traffic". In: *Proceedings of the 1998 IEEE International Conference on Control Applications (Cat. No. 98CH36104)*. Vol. 2. IEEE. 1998, pp. 868–872.

[5] Peng Jing et al. "The determinants behind the acceptance of autonomous vehicles: A systematic review". In: *Sustainability* 12.5 (2020), p. 1719.

[6] Christina Lichtenthäler et al. "Increasing perceived value between human and robots—measuring legibility in human aware navigation". In: *2012 IEEE workshop on advanced robotics and its social impacts (ARSO)*. IEEE. 2012, pp. 89–94.

[7] Kum Fai Yuen et al. "The determinants of public acceptance of autonomous vehicles: An innovation diffusion perspective". In: *Journal of Cleaner Production* 270 (2020), p. 121904.

[8] Matteo Rubagotti et al. "Perceived safety in physical human–robot interaction—A survey". In: *Robotics and Autonomous Systems* 151 (2022), p. 104047.

[9] Bentolhoda Jafary et al. "A survey on autonomous vehicles interactions with human and other vehicles". In: *14th PSAM International Conference on Probabilistic Safety Assessment and Management*. 2018.

[10] Olger Siebinga, Arkady Zgonnikov, and David Abbink. "A human factors approach to validating driver models for interaction-aware automated vehicles". In: *ACM Transactions on Human-Robot Interaction (THRI)* 11.4 (2022), pp. 1–21.

[11] Jeff X Ban. "A game theoretical approach for modelling merging and yielding behaviour at freeway on-ramp sections". In: *Transportation and Traffic Theory: Papers Selected for Presentation at 17th International Symposium on Transportation and Traffic Theory, a Peer Reviewed Series Since 1959*. Elsevier Amsterdam, The Netherlands. 2007, p. 197.

[12] Niclas Evestedt et al. "Interaction aware trajectory planning for merge scenarios in congested traffic situations". In: *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2016, pp. 465–472.

[13] Jaime F Fisac et al. “Hierarchical game-theoretic planning for autonomous vehicles”. In: *2019 International conference on robotics and automation (ICRA)*. IEEE. 2019, pp. 9590–9596.

[14] Xiangwang Hu and Jian Sun. “Trajectory optimization of connected and autonomous vehicles at a multilane freeway merging area”. In: *Transportation Research Part C: Emerging Technologies* 101 (2019), pp. 111–125.

[15] Chang Liu et al. “Path planning for autonomous vehicles using model predictive control”. In: *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2017, pp. 174–179.

[16] Wilko Schwarting et al. “Social behavior for autonomous vehicles”. In: *Proceedings of the National Academy of Sciences* 116.50 (2019), pp. 24972–24978.

[17] Meng Wang et al. “Game theoretic approach for predictive lane-changing and car-following control”. In: *Transportation Research Part C: Emerging Technologies* 58 (2015), pp. 73–92.

[18] Erik Ward et al. “Probabilistic model for interaction aware planning in merge scenarios”. In: *IEEE Transactions on Intelligent Vehicles* 2.2 (2017), pp. 133–146.

[19] Mofan Zhou, Xiaobo Qu, and Sheng Jin. “On the impact of cooperative autonomous vehicles in improving freeway merging: a modified intelligent driver model-based approach”. In: *IEEE Transactions on Intelligent Transportation Systems* 18.6 (2016), pp. 1422–1428.

[20] Shunsuke Aoki, Chung-Wei Lin, and Ragunathan Rajkumar. “Human-robot cooperation for autonomous vehicles and human drivers: Challenges and solutions”. In: *IEEE communications magazine* 59.8 (2021), pp. 35–41.

[21] John A Michon. “Explanatory pitfalls and rule-based driver models”. In: *Accident Analysis & Prevention* 21.4 (1989), pp. 341–353.

[22] Nicholas J Hetherington, Elizabeth A Croft, and HF Machiel Van der Loos. “Hey robot, which way are you going? nonverbal motion legibility cues for human-robot spatial interaction”. In: *IEEE Robotics and Automation Letters* 6.3 (2021), pp. 5010–5015.

[23] Dongfang Niu, Jacques Terken, and Berry Eggen. “Anthropomorphizing information to enhance trust in autonomous vehicles”. In: *Human Factors and Ergonomics in Manufacturing & Service Industries* 28.6 (2018), pp. 352–359.

[24] Alexey Dosovitskiy et al. “CARLA: An Open Urban Driving Simulator”. In: *Proceedings of the 1st Annual Conference on Robot Learning*. 2017, pp. 1–16.

[25] Niek Beckers et al. “JOAN: a framework for human-automated vehicle interaction experiments in a virtual reality driving simulator”. In: *Journal of Open Source Software* 8.82 (2023), p. 4250.

[26] Harold R Booher. *MANPRINT: An approach to systems integration*. Springer Science & Business Media, 2012.

[27] Aaron Steinfeld et al. “Common metrics for human-robot interaction”. In: *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 2006, pp. 33–40.

[28] Bernard CK Choi and Anita WP Pak. “Peer reviewed: a catalog of biases in questionnaires”. In: *Preventing chronic disease* 2.1 (2005).

[29] George D Ogden, Jerrold M Levine, and Ellen J Eisner. “Measurement of workload by secondary tasks”. In: *Human Factors* 21.5 (1979), pp. 529–548.

[30] PGAM Jorna. “Heart rate and workload variations in actual and simulated flight”. In: *Ergonomics* 36.9 (1993), pp. 1043–1054.

[31] Anne-Marie Brouwer et al. “Estimating workload using EEG spectral power and ERPs in the n-back task”. In: *Journal of neural engineering* 9.4 (2012), p. 045008.

[32] JA Veltman and AWK Gaillard. “Physiological workload reactions to increasing levels of task difficulty”. In: *Ergonomics* 41.5 (1998), pp. 656–669.

[33] Joost CF De Winter et al. “Effects of adaptive cruise control and highly automated driving on workload and situation awareness: A review of the empirical evidence”. In: *Transportation research part F: traffic psychology and behaviour* 27 (2014), pp. 196–217.

[34] Jung-Chun Liu et al. “Assessing perceptual load and cognitive load by fixation-related information of eye movements”. In: *Sensors* 22.3 (2022), p. 1187.

[35] Emma J Nilsson et al. “On-to-off-path gaze shift cancellations lead to gaze concentration in cognitively loaded car drivers: A simulator study exploring gaze patterns in relation to a cognitive task and the traffic environment”. In: *Transportation research part F: traffic psychology and behaviour* 75 (2020), pp. 1–15.

[36] Vérane Faure, Régis Lobjois, and Nicolas Benguigui. “The effects of driving environment complexity and dual tasking on drivers’ mental workload and eye blink behavior”. In: *Transportation research part F: traffic psychology and behaviour* 40 (2016), pp. 78–90.

[37] Simone Benedetto et al. “Driver workload and eye blink duration”. In: *Transportation research part F: traffic psychology and behaviour* 14.3 (2011), pp. 199–208.

[38] Nadine Matton, Pierre-Vincent Paubel, and Sébastien Puma. “Toward the use of pupillary responses for pilot selection”. In: *Human factors* 64.3 (2022), pp. 555–567.

[39] Ganesh Pai Mangalore et al. “The promise of virtual reality headsets: Can they be used to measure accurately drivers’ hazard anticipation performance?” In: *Transportation research record* 2673.10 (2019), pp. 455–464.

[40] Christoph Bartneck et al. “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots”. In: *International journal of social robotics* 1 (2009), pp. 71–81.

[41] Colleen M Carpinella et al. “The robotic social attributes scale (RoSAS) development and validation”. In: *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 2017, pp. 254–262.

[42] Michiel Joosse et al. “BEHAVE-II: The revised set of measures to assess users’ attitudinal and behavioral responses to a social robot”. In: *International journal of social robotics* 5 (2013), pp. 379–388.

[43] Jun Ma et al. “A systematic review on the influence factors, measurement, and effect of driver workload”. In: *Accident Analysis & Prevention* 192 (2023), p. 107289.

[44] Weiya Chen, Tetsuo Sawaragi, and Toshihiro Hiraoka. “Comparing eye-tracking metrics of mental workload

caused by NDRTs in semi-autonomous driving". In: *Transportation research part F: traffic psychology and behaviour* 89 (2022), pp. 109–128.

[45] Weiya Chen, Tetsuo Sawaragi, and Yukio Horiguchi. "Measurement of driver's mental workload in partial autonomous driving". In: *IFAC-PapersOnLine* 52.19 (2019), pp. 347–352.

[46] Jork Stapel, Freddy Antony Mullakkal-Babu, and Riender Happee. "Automated driving reduces perceived workload, but monitoring causes higher cognitive load than manual driving". In: *Transportation research part F: traffic psychology and behaviour* 60 (2019), pp. 590–605.

[47] Jean Scholtz. "Theory and evaluation of human robot interactions". In: *36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the*. IEEE. 2003, 10–pp.

[48] Mica R Endsley. "Situation awareness in aircraft systems: Symposium abstract". In: *Proceedings of the Human Factors Society Annual Meeting*. Vol. 32. 2. SAGE Publications Sage CA: Los Angeles, CA. 1988, pp. 96–96.

[49] Federico Scari. "Assessment of Autonomous Vehicle Controllers based on the Interaction with Human-driven Vehicles". In: (2023).

[50] Geoffrey Underwood, David Crundall, and Peter Chapman. "Driving simulator validation with hazard perception". In: *Transportation research part F: traffic psychology and behaviour* 14.6 (2011), pp. 435–446.

APPENDIX A

Appendix A shows the average lateral and longitudinal positions of the vehicles during experiment 1 divided by conditions Figure 63, the boxplots for the fixation frequency for experiment 1 Figure 14, the boxplots for the total duration of the fixation for experiment 1 Figure 15, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 1 Figure 16 and the Perceived Safety during experiment 1 Figure 17.

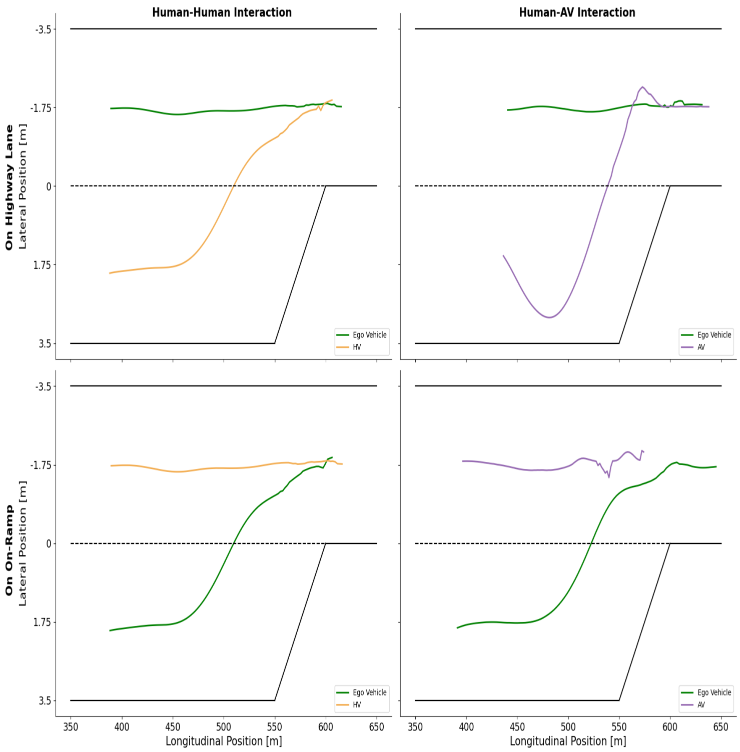

Figure 13: Average Lateral and longitudinal positions of the vehicles during experiment 1

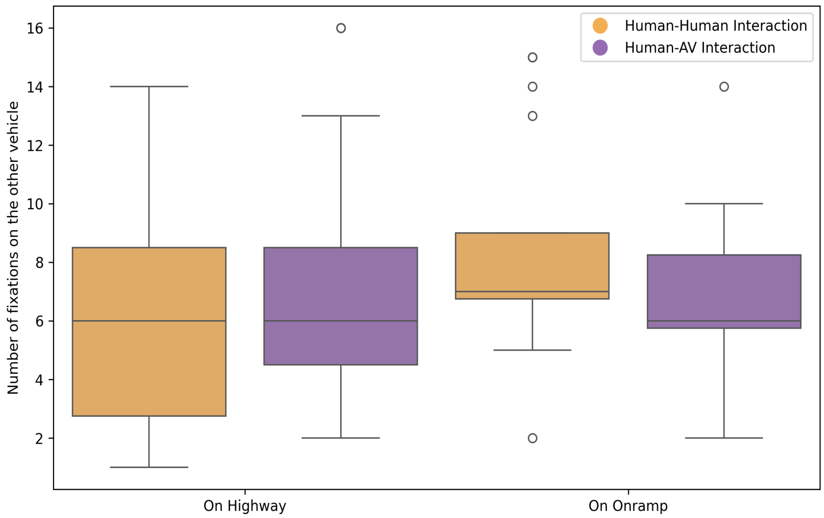

Figure 14: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

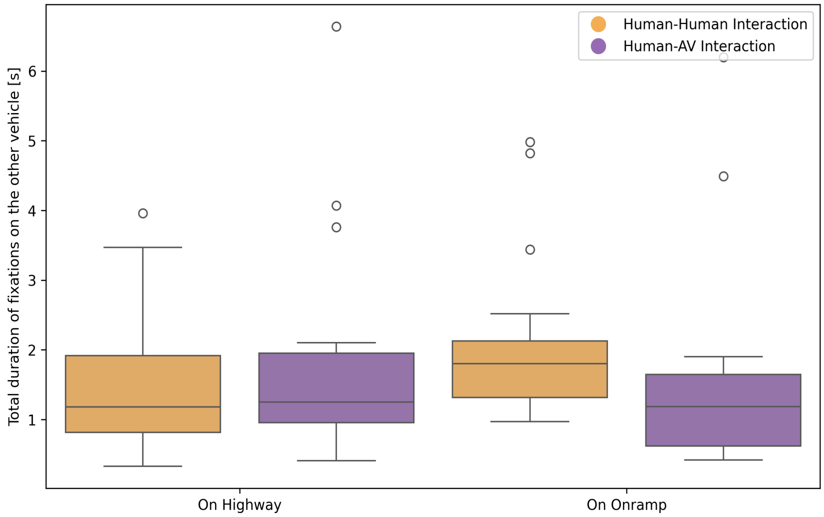

Figure 15: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

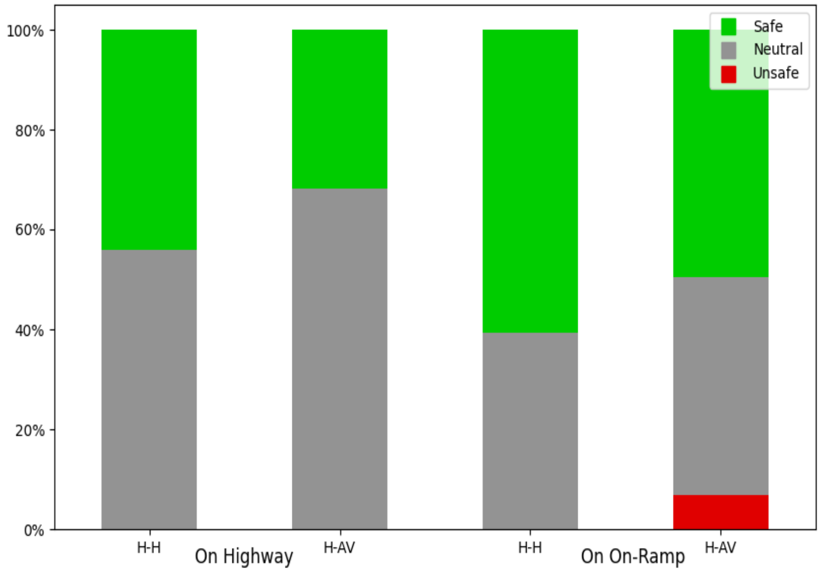

Figure 16: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

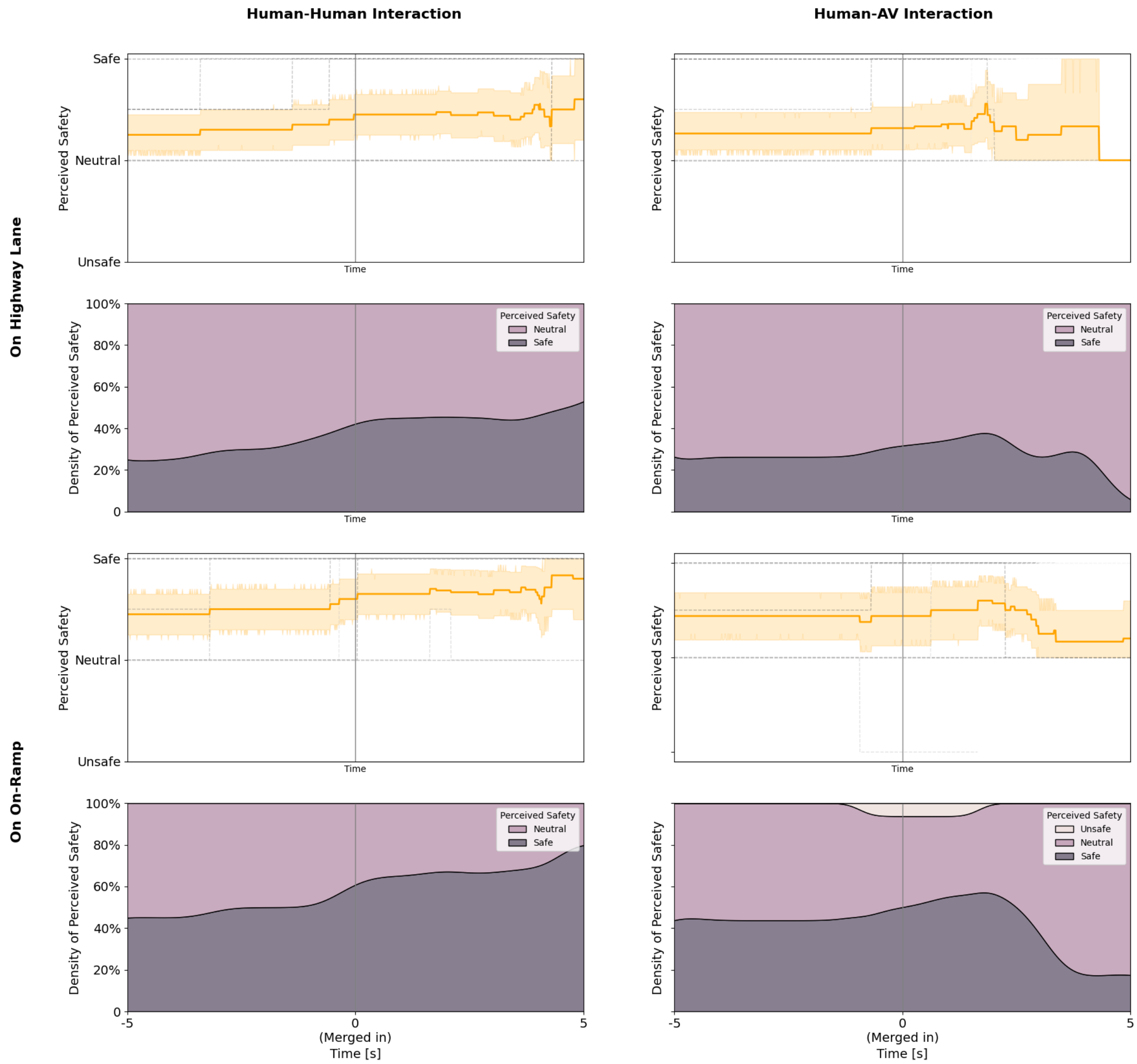

Figure 17: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

Appendix B

Appendix B shows the average lateral and longitudinal positions of the vehicles during experiment 2 divided by conditions Figure 18, the boxplots for the fixation frequency for experiment 2 Figure 19, the boxplots for the total duration of the fixation for experiment 2 Figure 20, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 2 Figure 21 and the Perceived Safety during experiment 2 Figure 22.

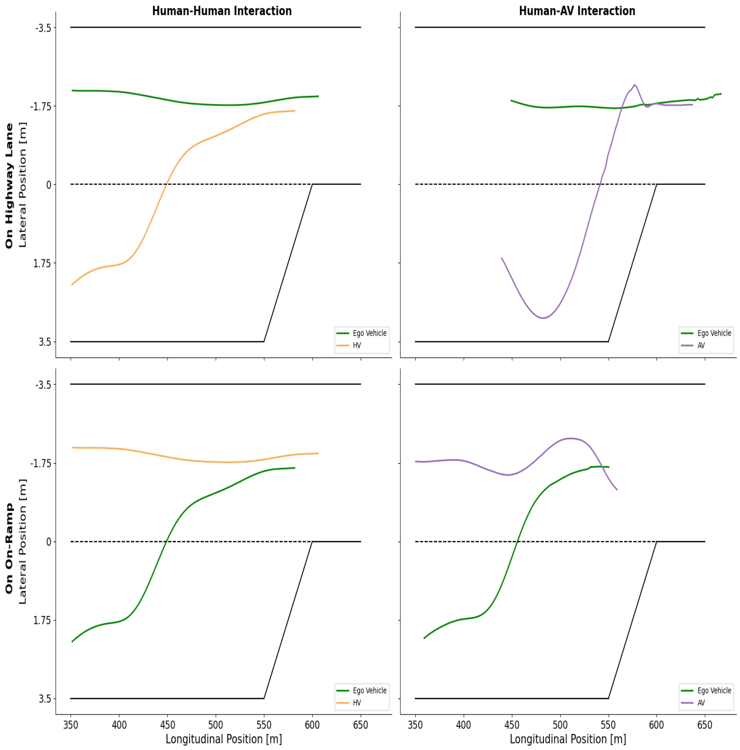

Figure 18: Average Lateral and longitudinal positions of the vehicles during experiment 1

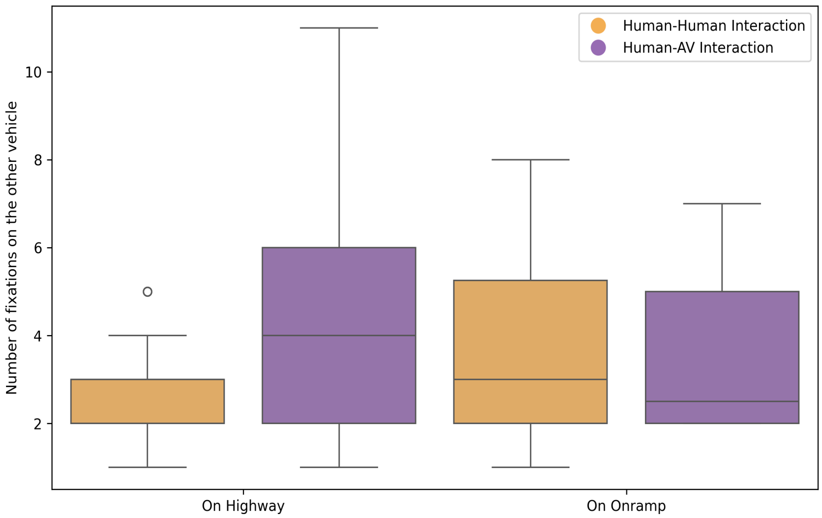

Figure 19: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

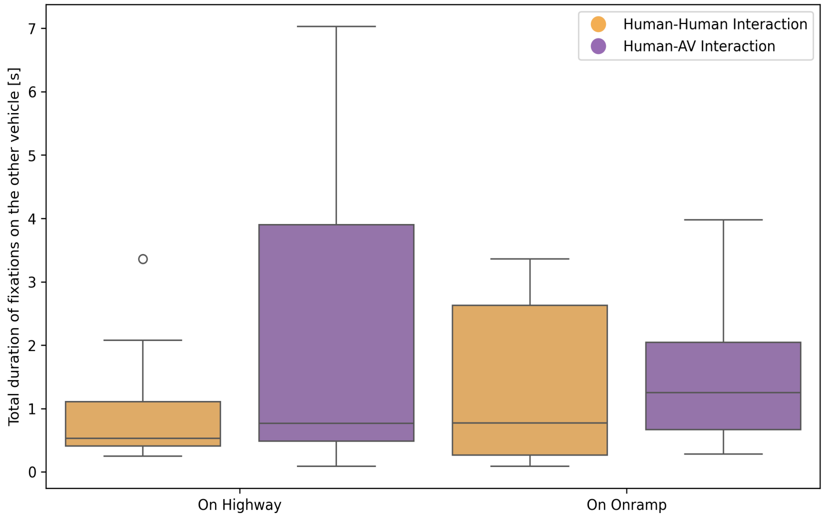

Figure 20: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

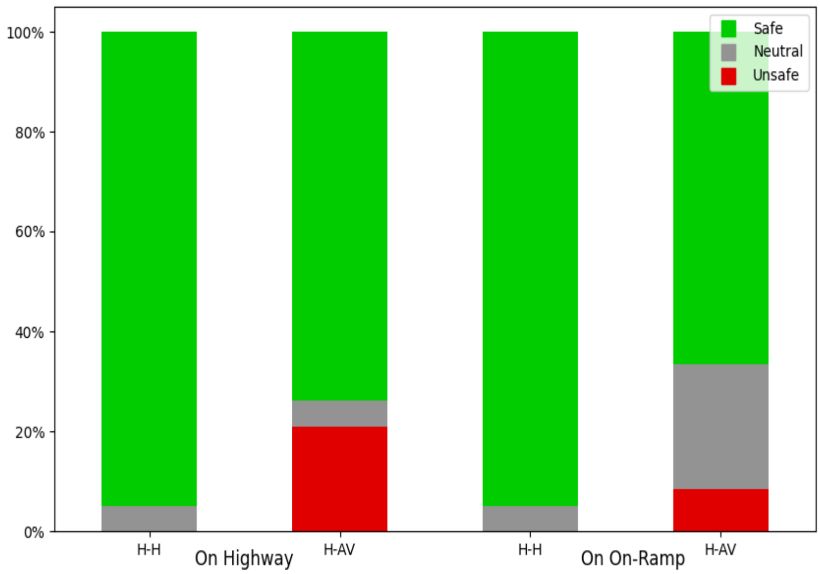

Figure 21: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

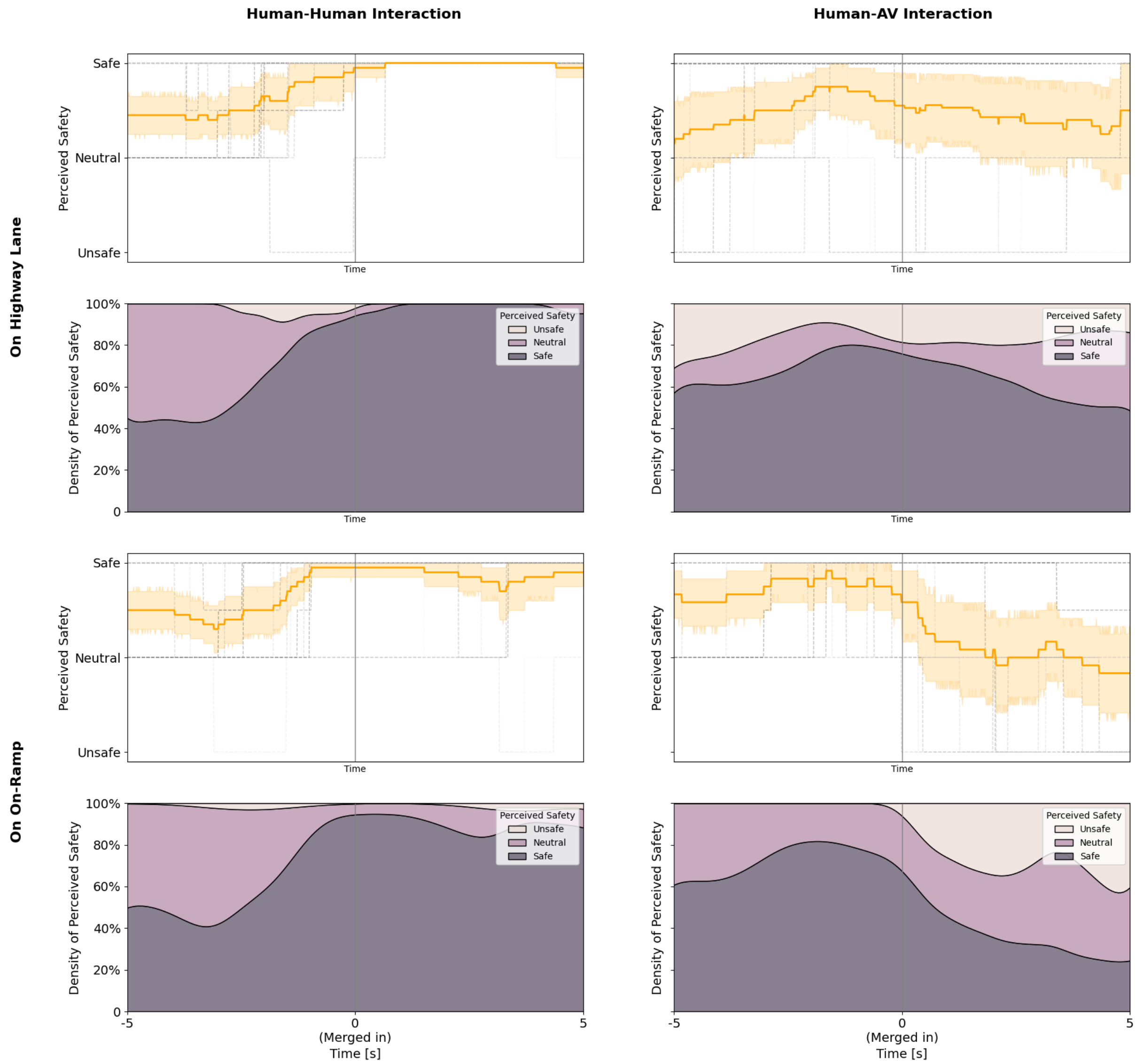

Figure 22: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

APPENDIX C

Appendix C shows the average lateral and longitudinal positions of the vehicles during experiment 3 divided by conditions Figure 23, the boxplots for the fixation frequency for experiment 3 Figure 24, the boxplots for the total duration of the fixation for experiment 3 Figure 25, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 3 Figure 26 and the Perceived Safety during experiment 3 Figure 27.

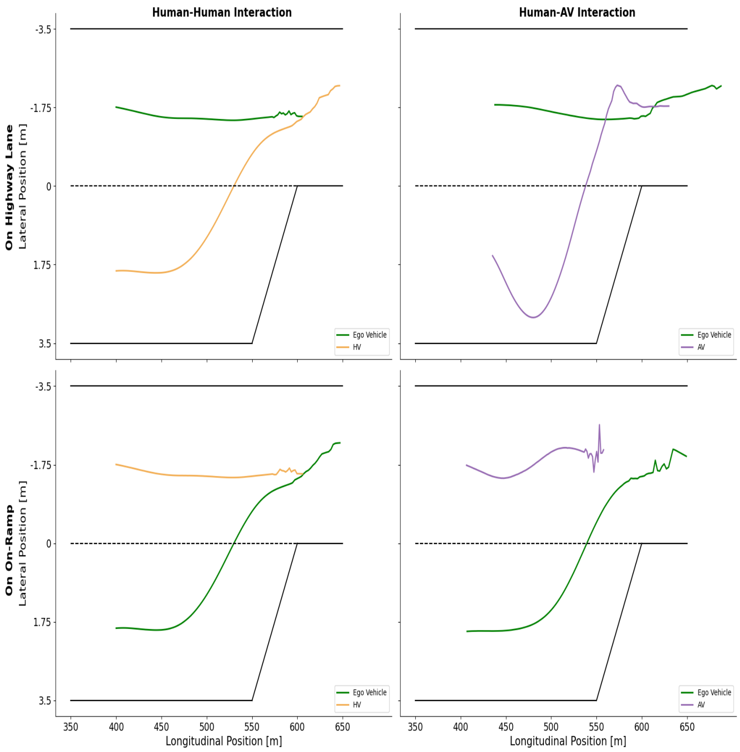

Figure 23: Average Lateral and longitudinal positions of the vehicles during experiment 1

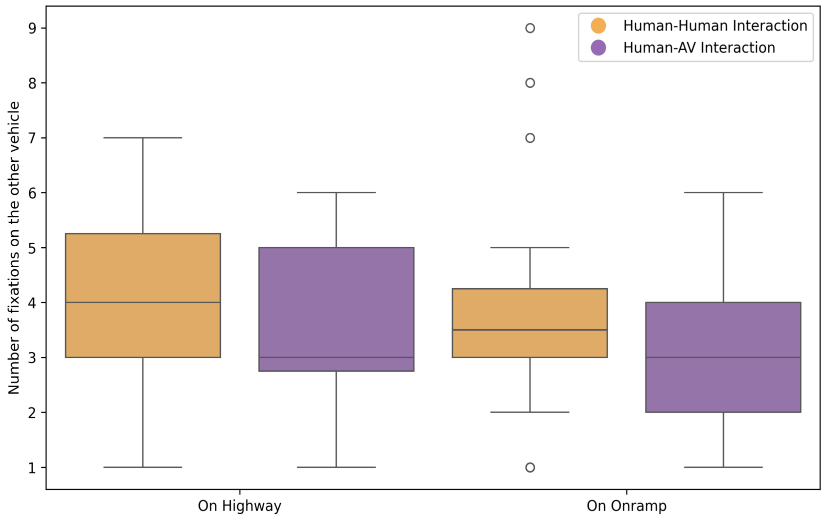

Figure 24: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

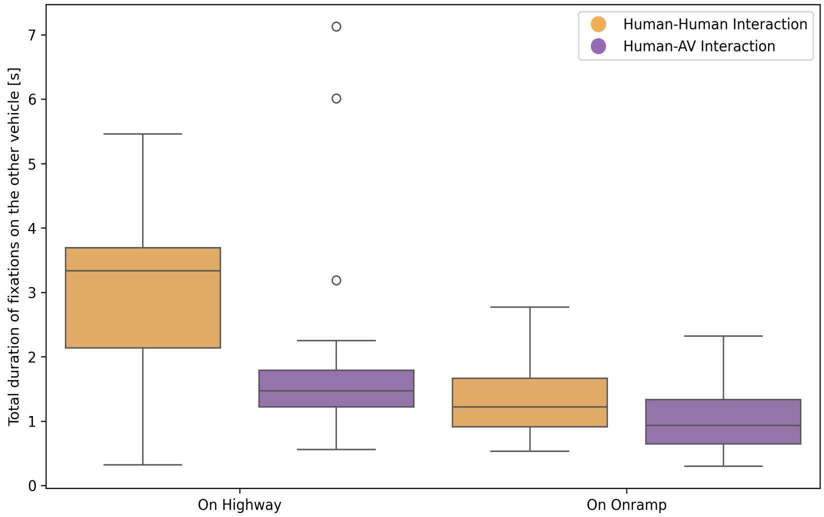

Figure 25: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

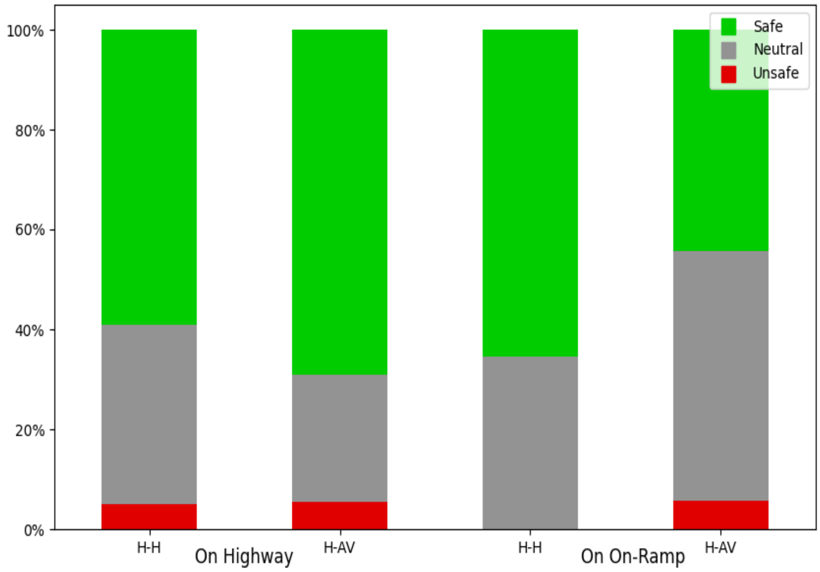

Figure 26: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

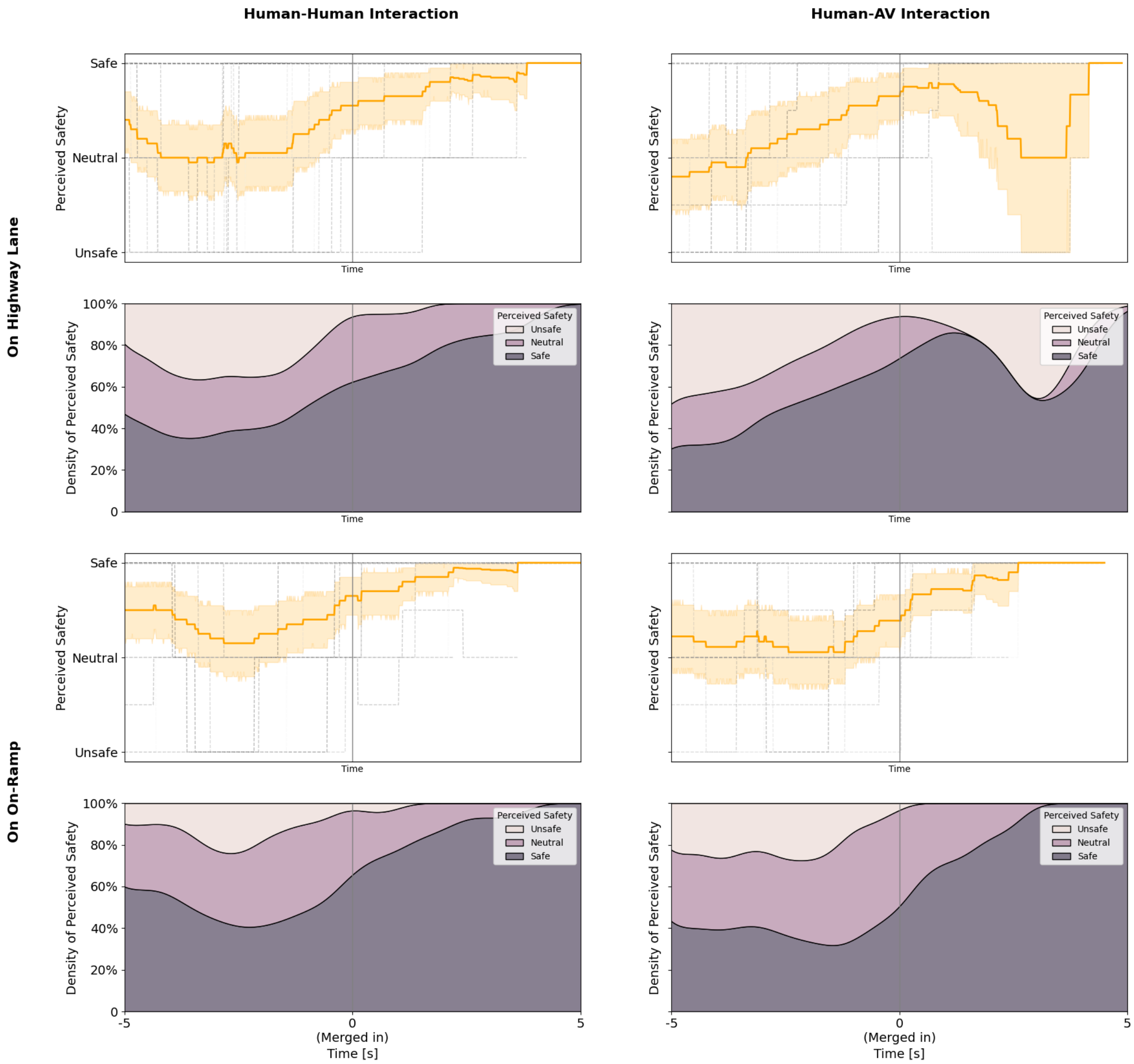

Figure 27: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

Appendix D

Appendix D shows the average lateral and longitudinal positions of the vehicles during experiment 4 divided by conditions Figure 28, the boxplots for the fixation frequency for experiment 4 Figure 29, the boxplots for the total duration of the fixation for experiment 4 Figure 30, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 4 Figure 31 and the Perceived Safety during experiment 4 Figure 32.

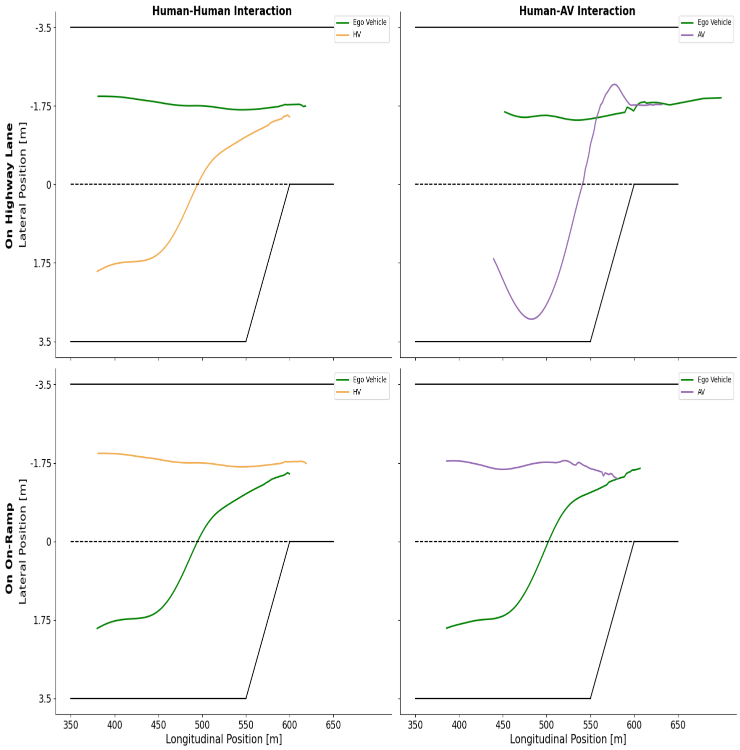

Figure 28: Average Lateral and longitudinal positions of the vehicles during experiment 1

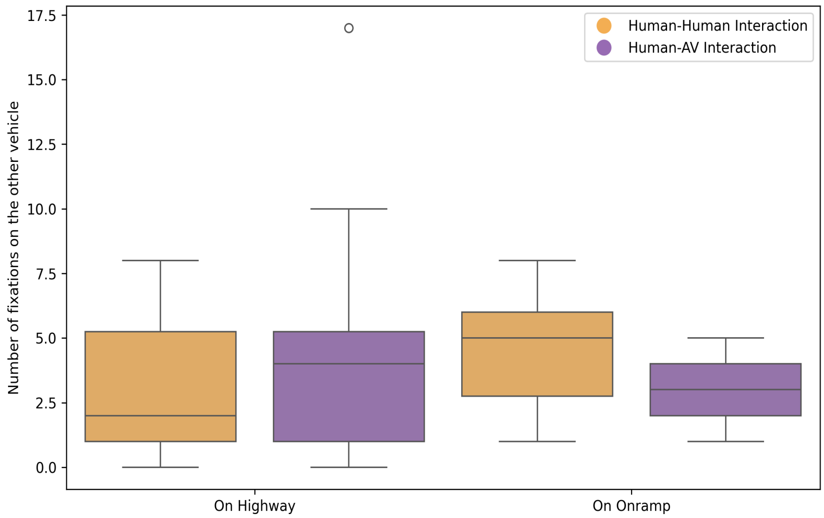

Figure 29: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

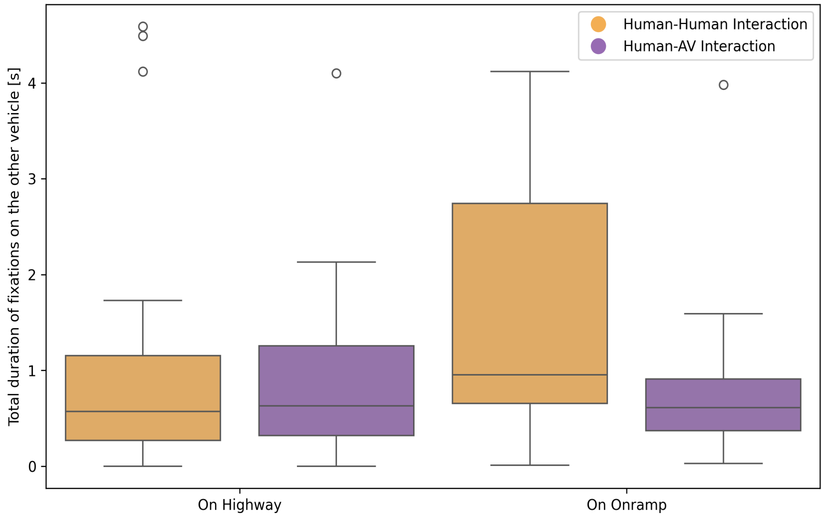

Figure 30: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

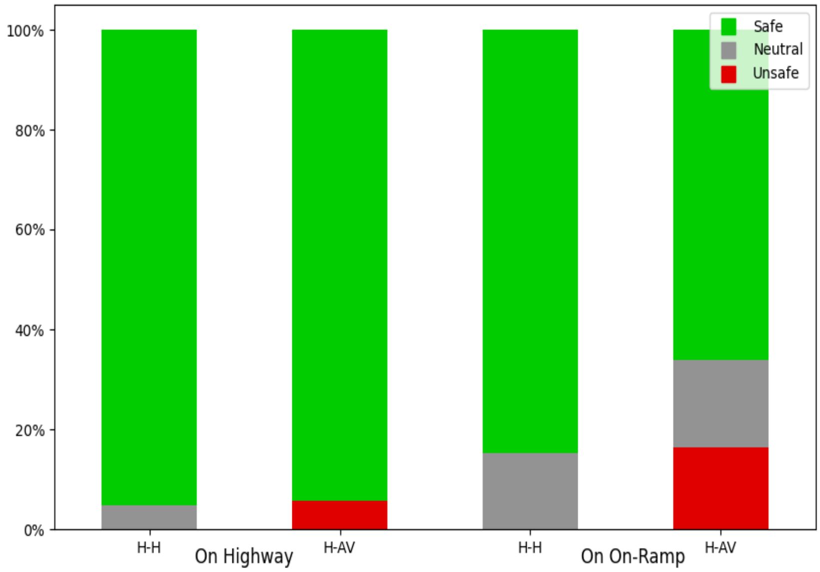

Figure 31: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

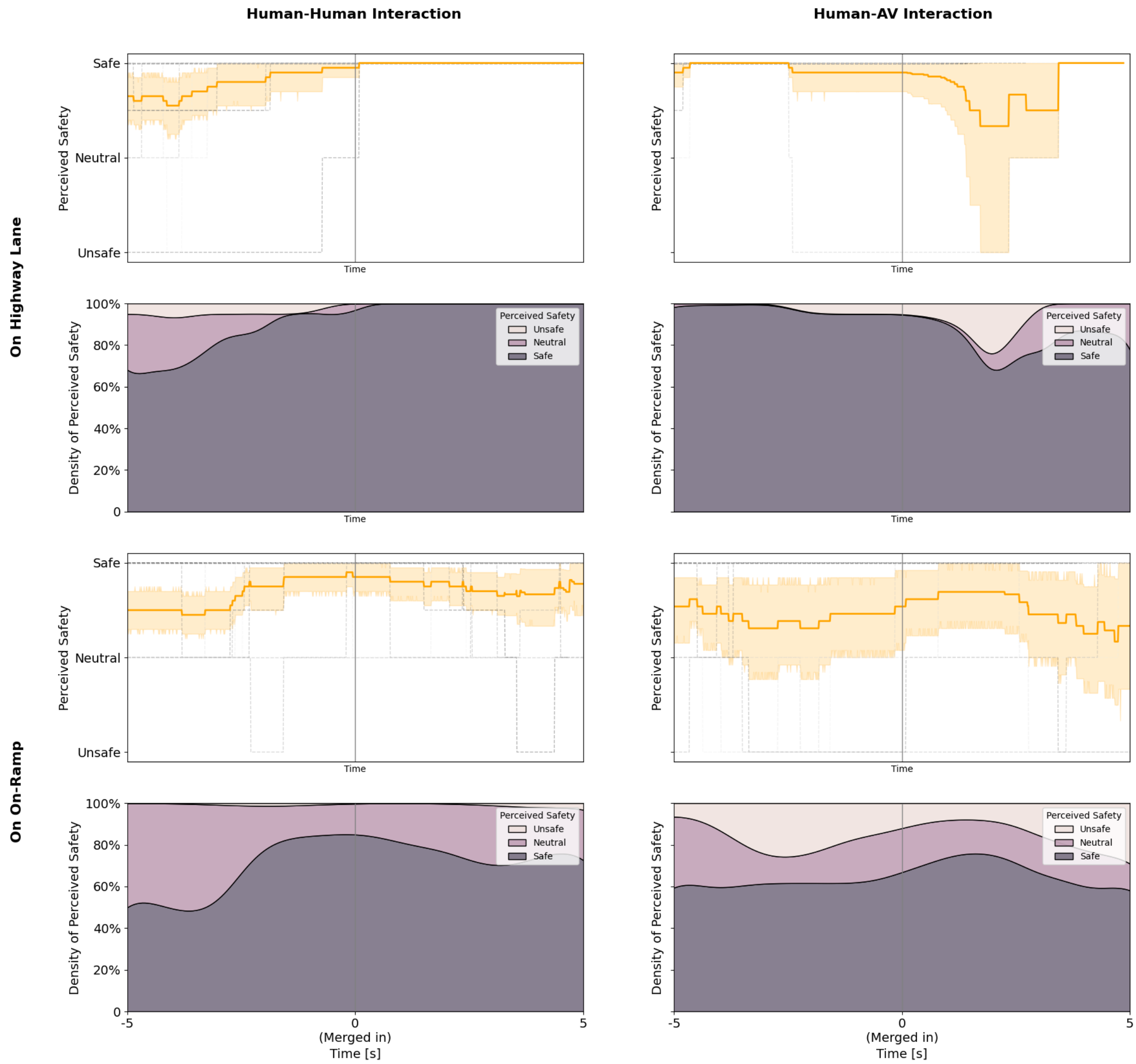

Figure 32: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

APPENDIX E

Appendix E shows the average lateral and longitudinal positions of the vehicles during experiment 5 divided by conditions Figure 33, the boxplots for the fixation frequency for experiment 5 Figure 34, the boxplots for the total duration of the fixation for experiment 5 Figure 35, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 5 Figure 36 and the Perceived Safety during experiment 5 Figure 37.

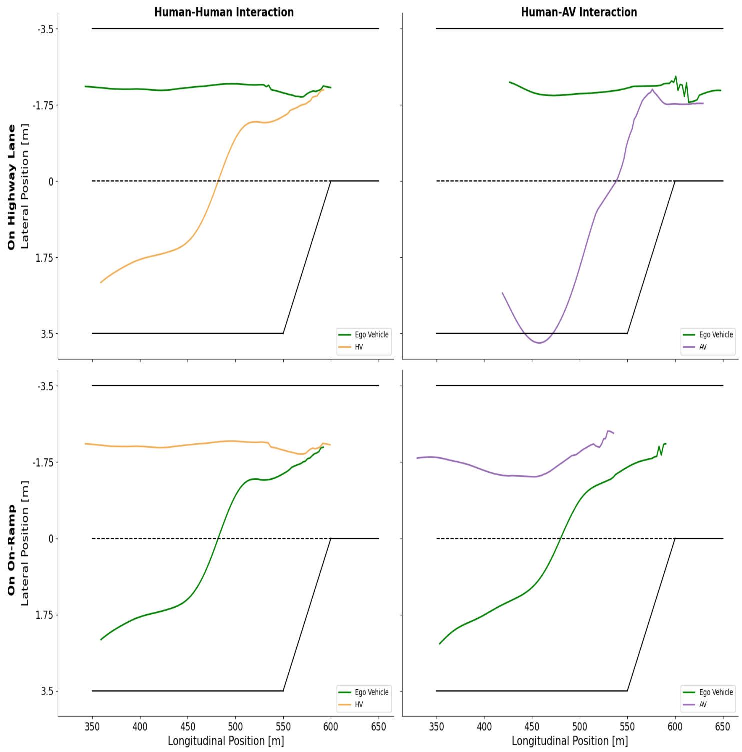

Figure 33: Average Lateral and longitudinal positions of the vehicles during experiment 1

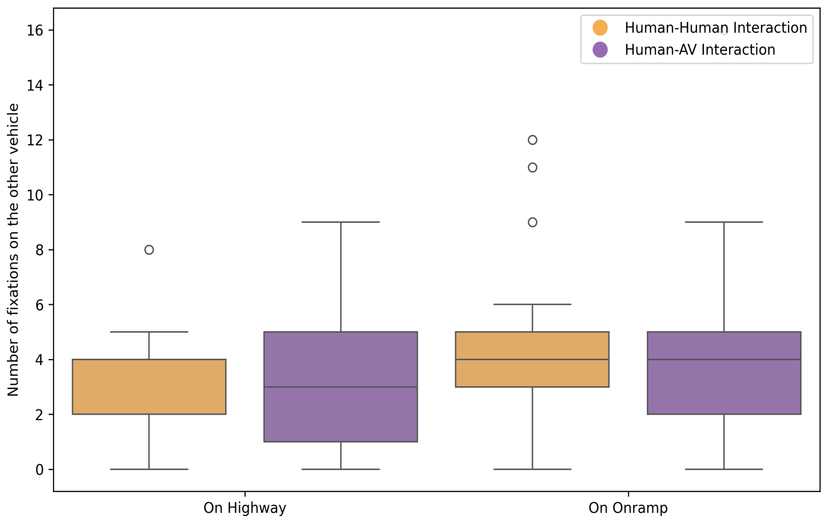

Figure 34: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

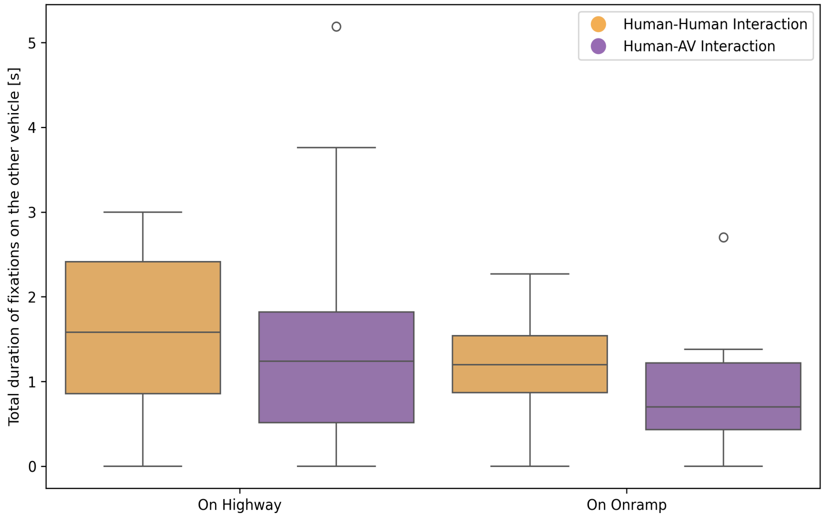

Figure 35: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

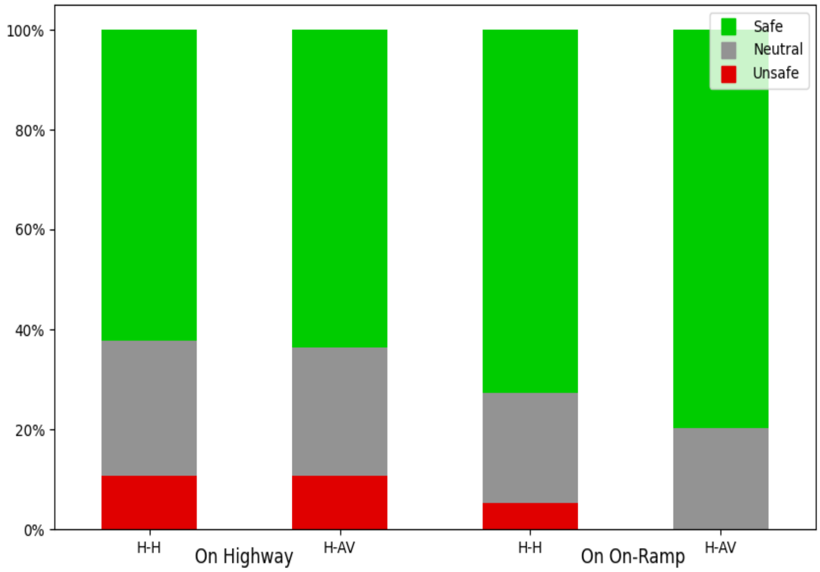

Figure 36: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

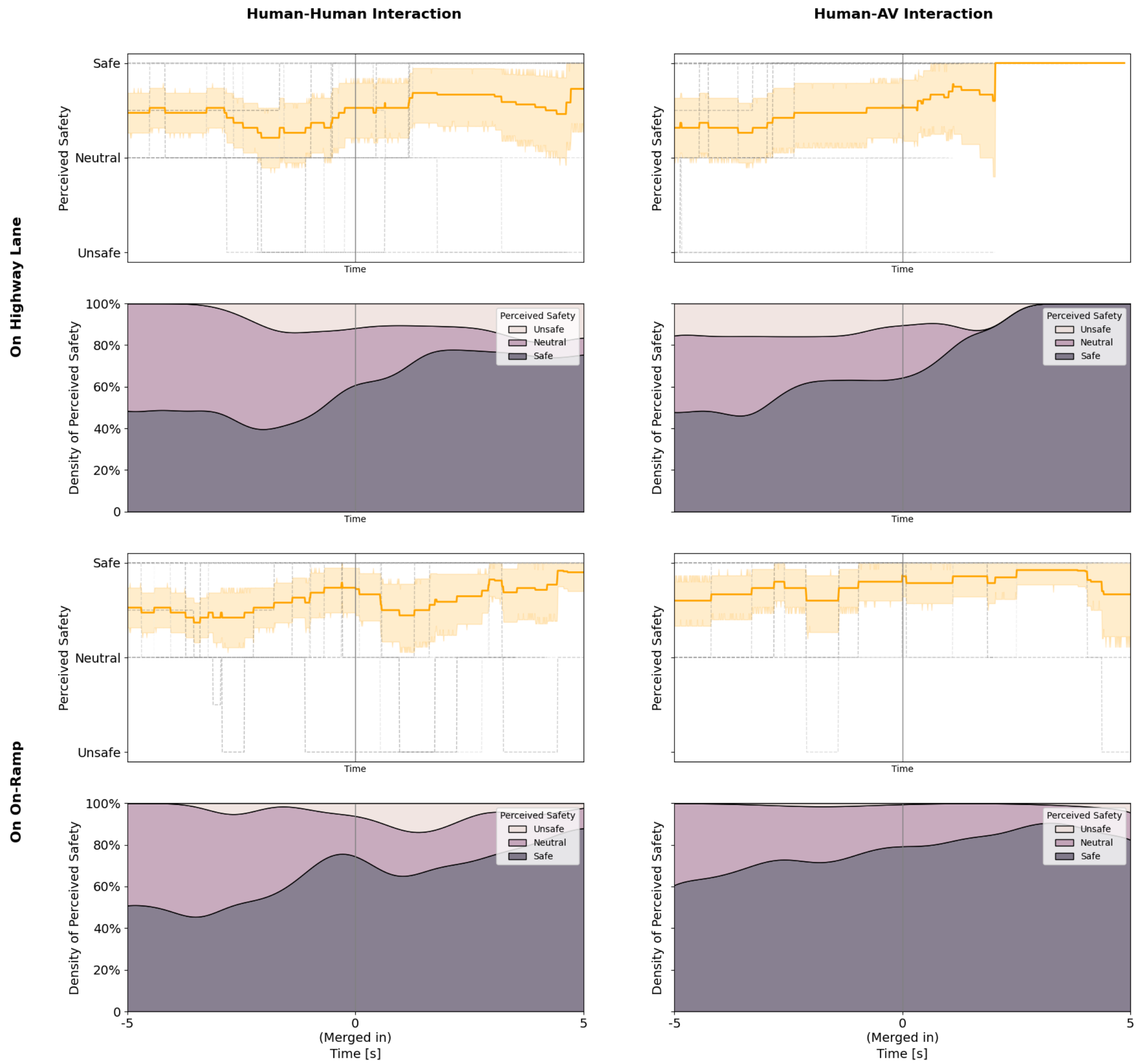

Figure 37: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

APPENDIX F

Appendix F shows the average lateral and longitudinal positions of the vehicles during experiment 6 divided by conditions Figure 38, the boxplots for the fixation frequency for experiment 6 Figure 39, the boxplots for the total duration of the fixation for experiment 6 Figure 40, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 6 Figure 41 and the Perceived Safety during experiment 6 Figure 42.

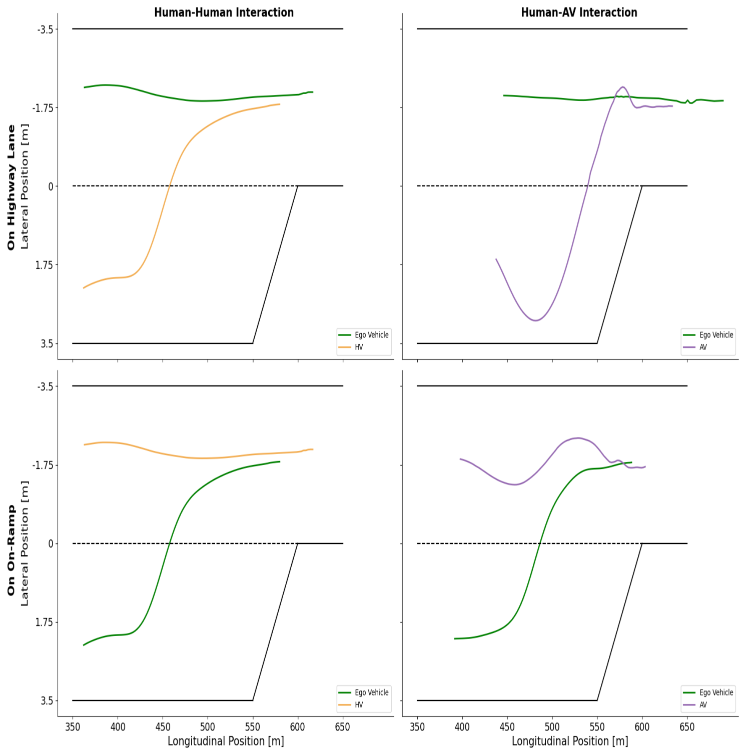

Figure 38: Average Lateral and longitudinal positions of the vehicles during experiment 1

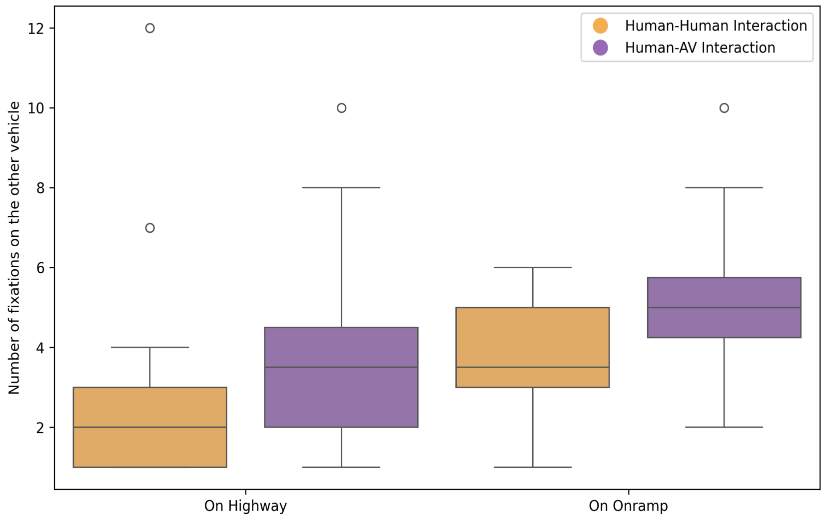

Figure 39: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

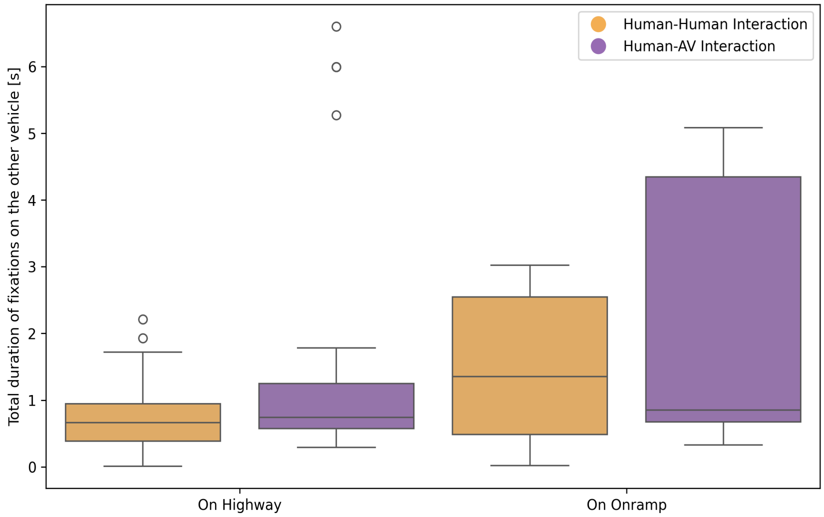

Figure 40: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

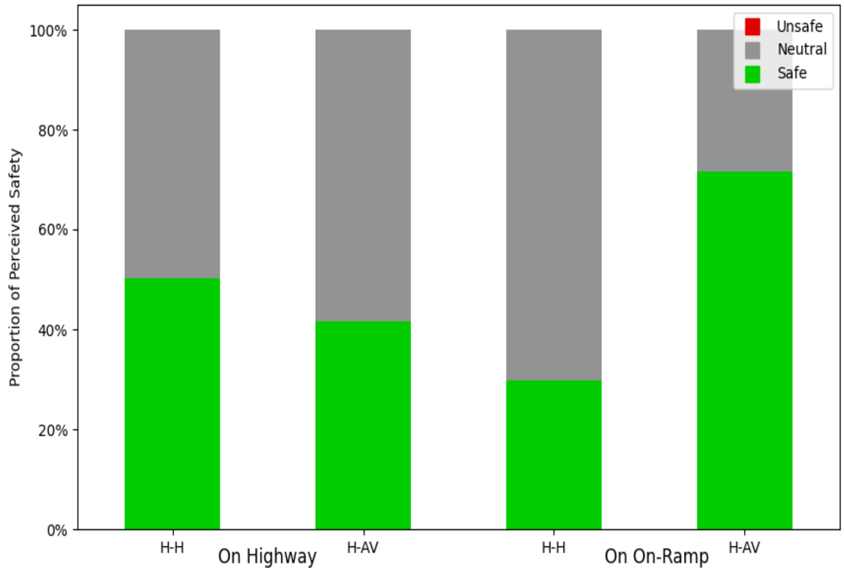

Figure 41: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

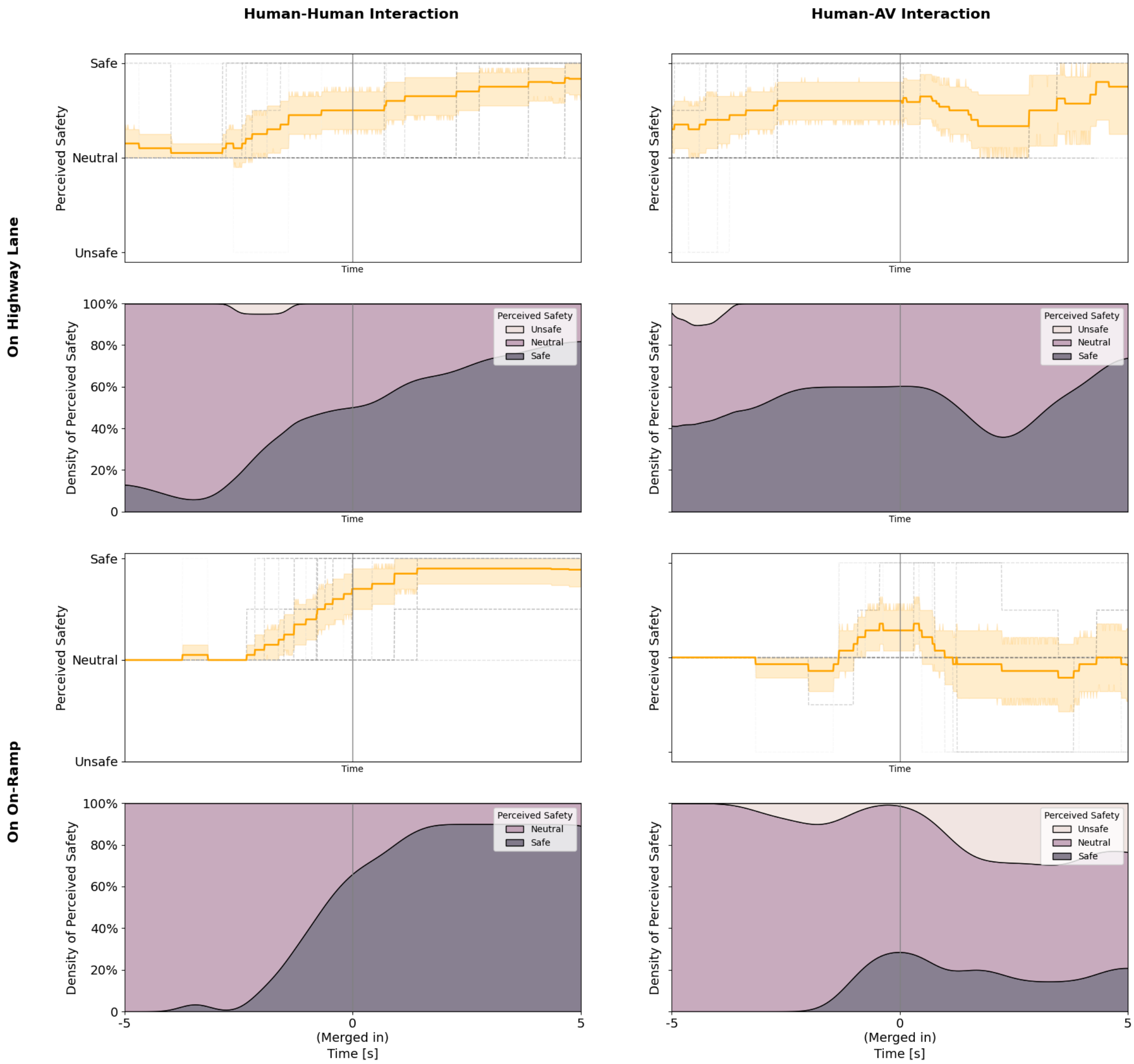

Figure 42: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

APPENDIX G

Appendix G shows the average lateral and longitudinal positions of the vehicles during experiment 7 divided by conditions Figure 43, the boxplots for the fixation frequency for experiment 7 Figure 44, the boxplots for the total duration of the fixation for experiment 7 Figure 45, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 7 Figure 46 and the Perceived Safety during experiment 7 Figure 47.

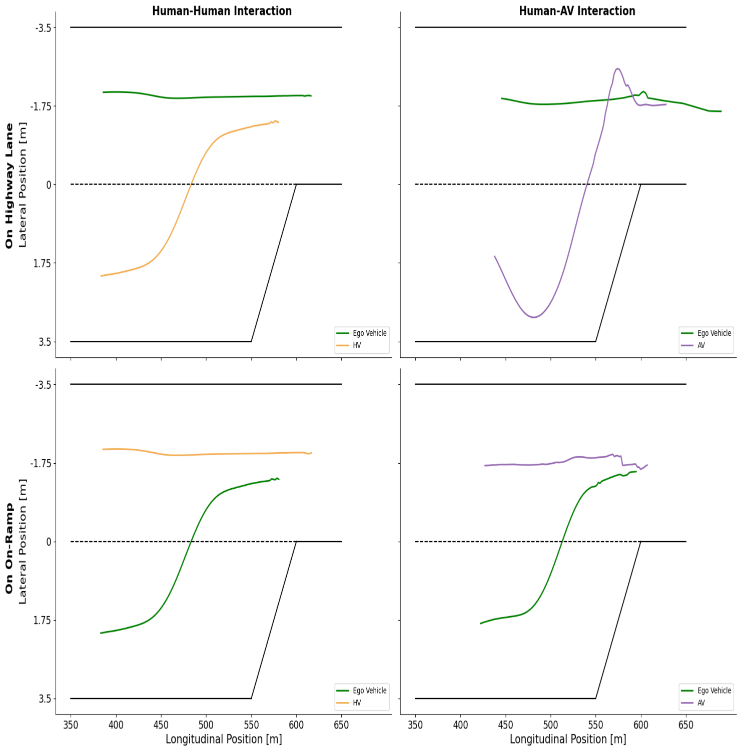

Figure 43: Average Lateral and longitudinal positions of the vehicles during experiment 1

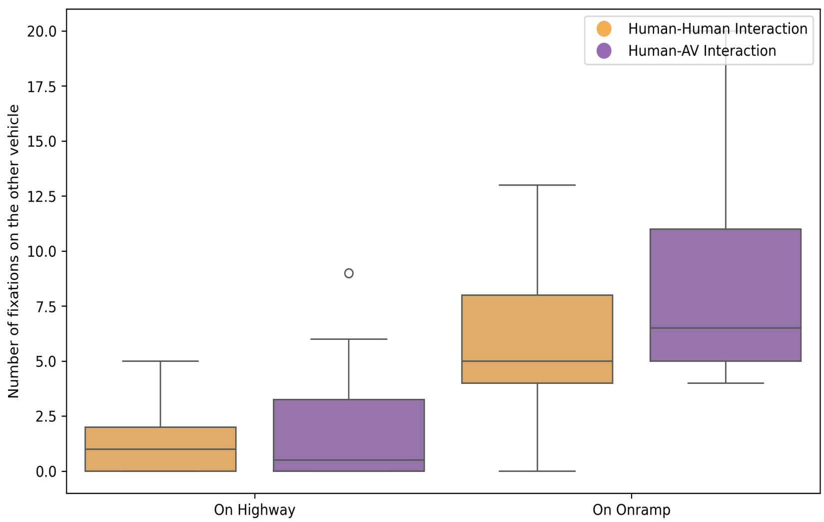

Figure 44: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

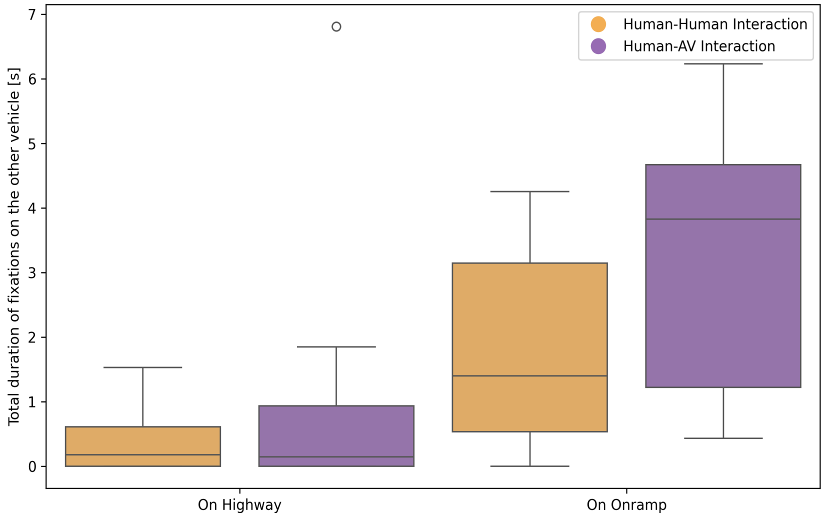

Figure 45: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

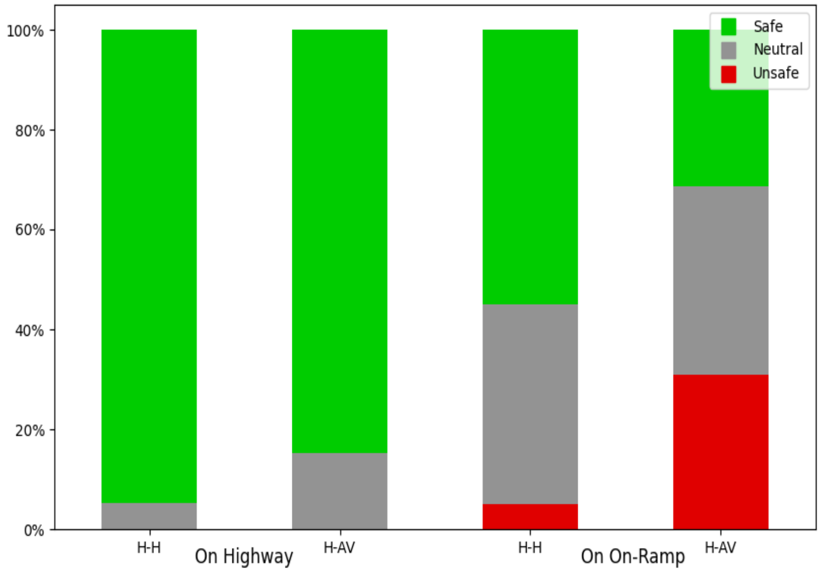

Figure 46: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

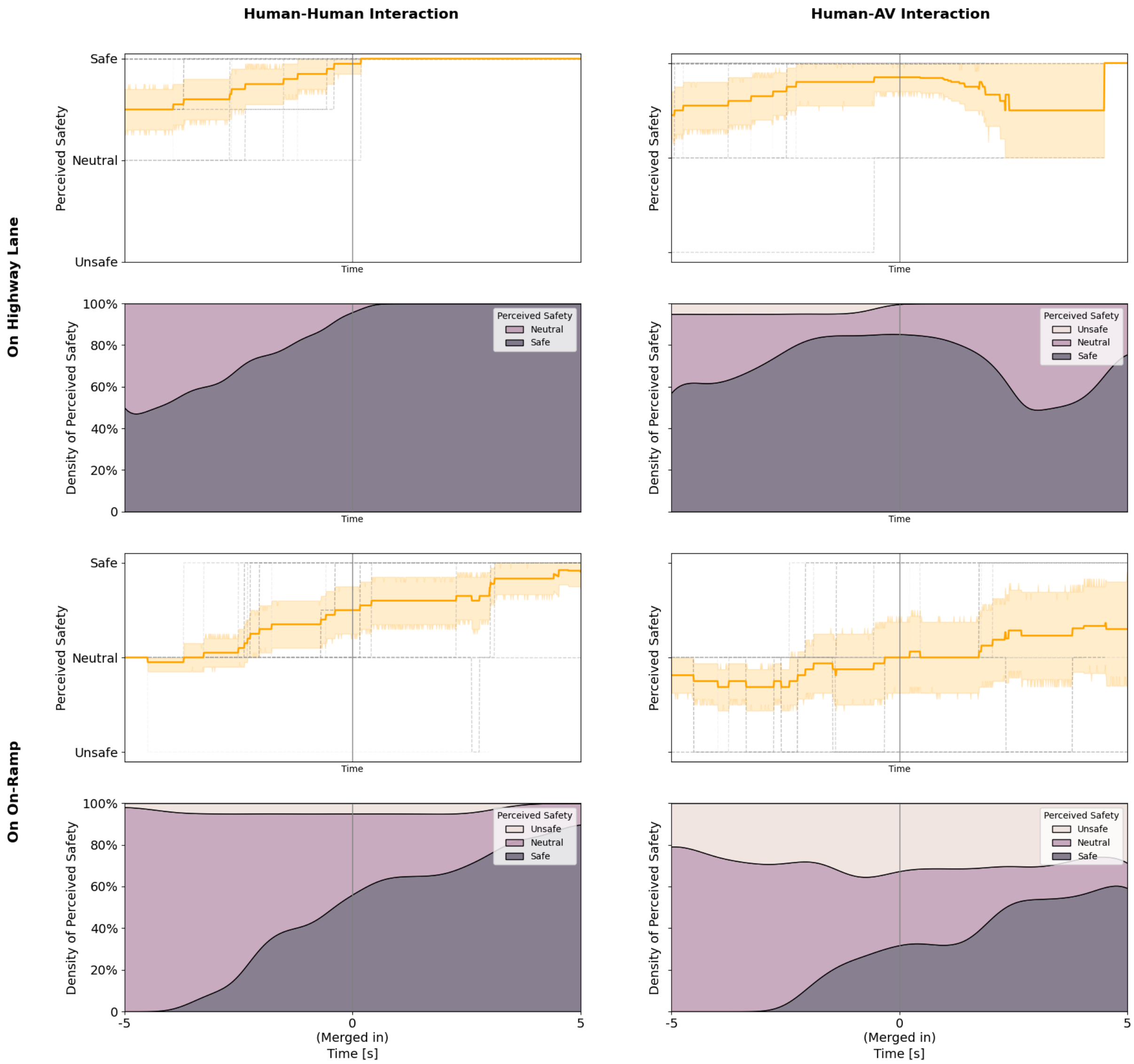

Figure 47: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

Appendix H

Appendix H shows the average lateral and longitudinal positions of the vehicles during experiment 8 divided by conditions Figure 48, the boxplots for the fixation frequency for experiment 8 Figure 49, the boxplots for the total duration of the fixation for experiment 8 Figure 50, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 8 Figure 51 and the Perceived Safety during experiment 8 Figure 52.

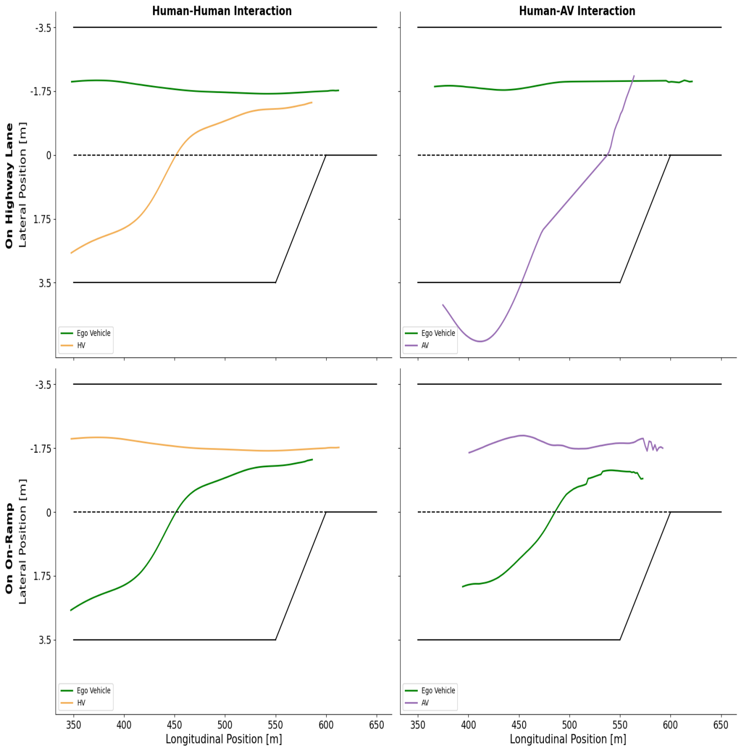

Figure 48: Average Lateral and longitudinal positions of the vehicles during experiment 1

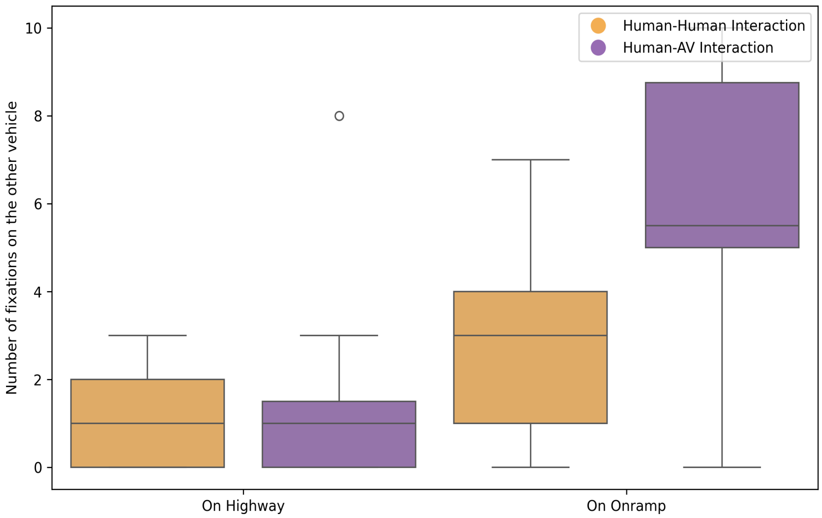

Figure 49: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

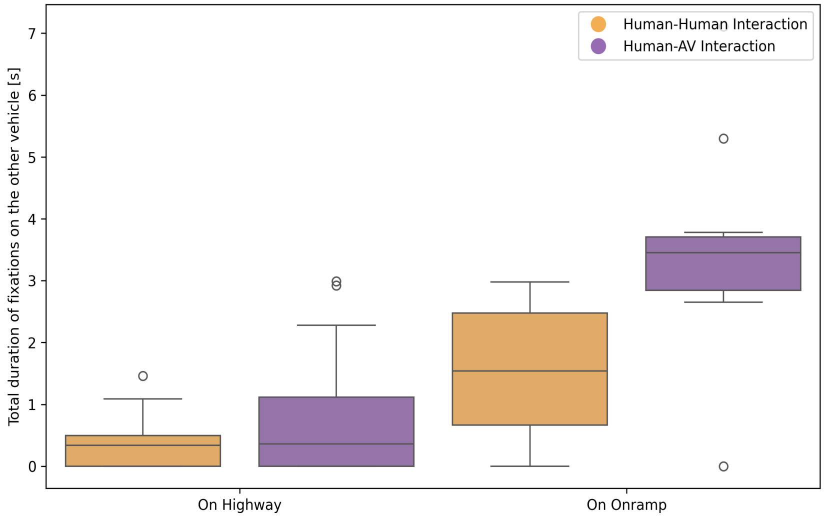

Figure 50: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

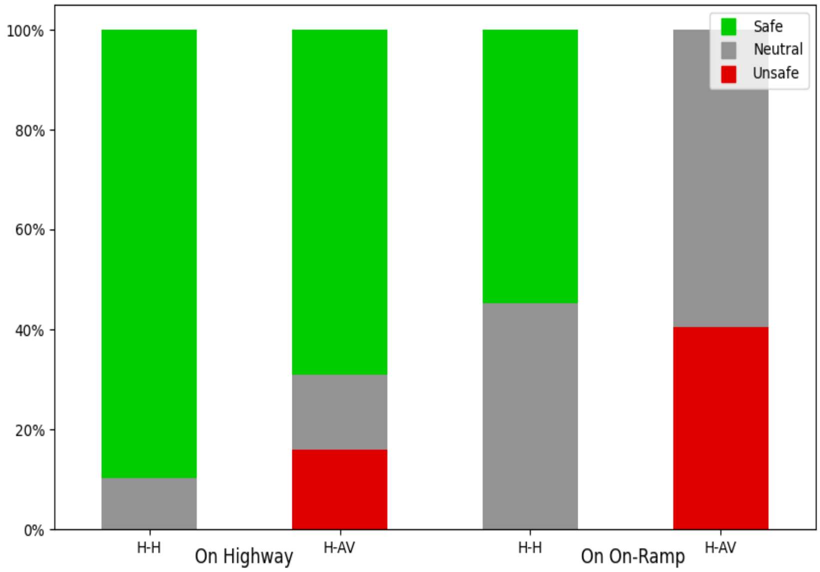

Figure 51: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

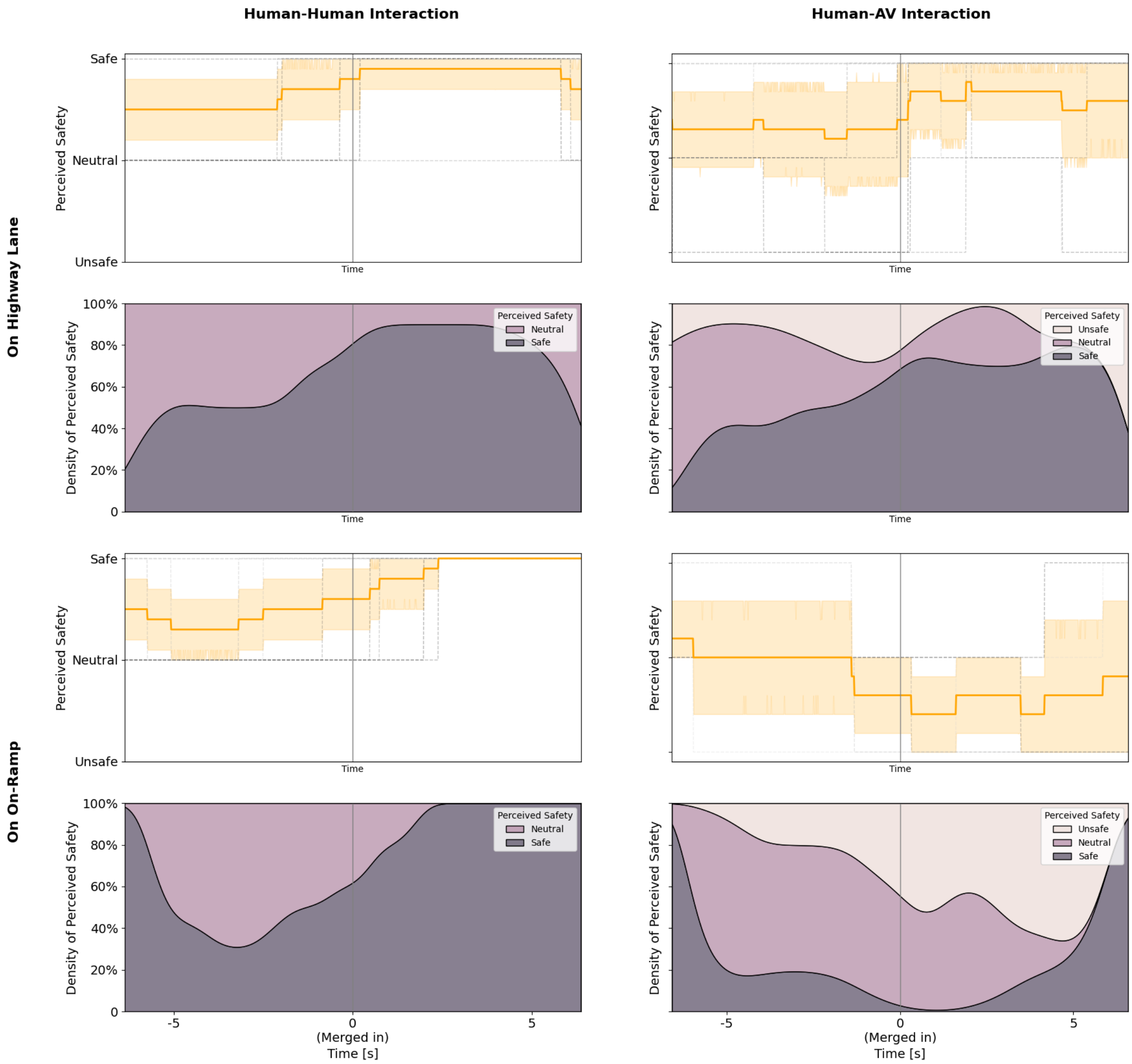

Figure 52: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

Appendix I

Appendix I shows the average lateral and longitudinal positions of the vehicles during experiment 9 divided by conditions Figure 53, the boxplots for the fixation frequency for experiment 9 Figure 54, the boxplots for the total duration of the fixation for experiment 9 Figure 55, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 9 Figure 56 and the Perceived Safety during experiment 9 Figure 57.

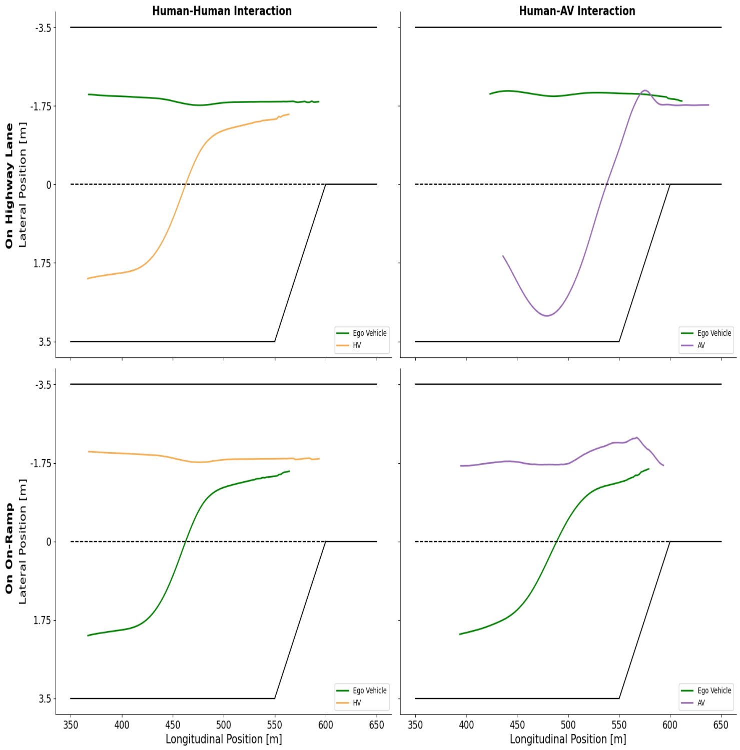

Figure 53: Average Lateral and longitudinal positions of the vehicles during experiment 1

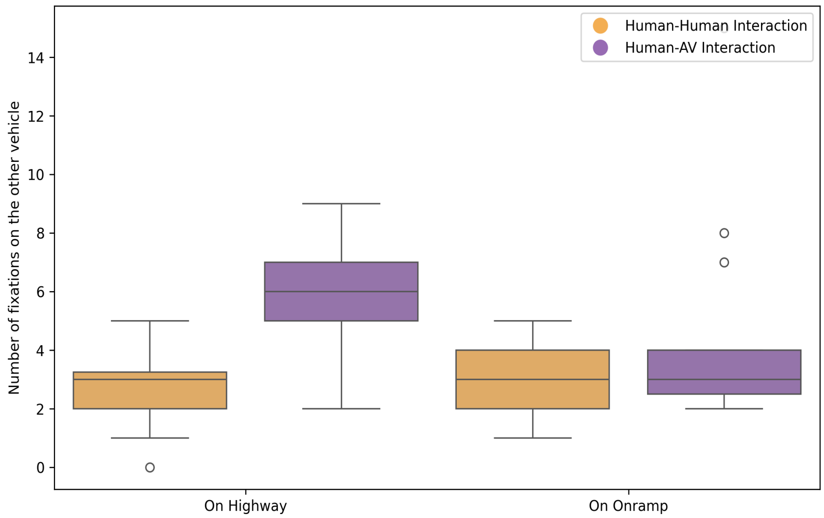

Figure 54: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

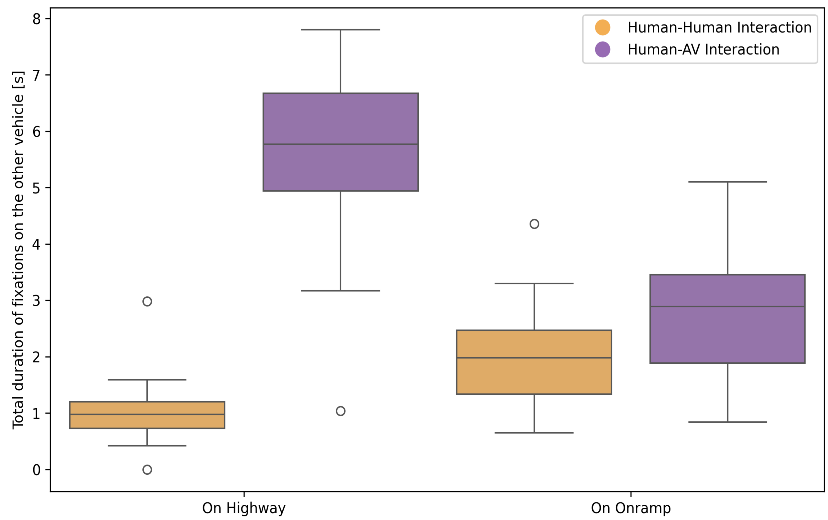

Figure 55: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

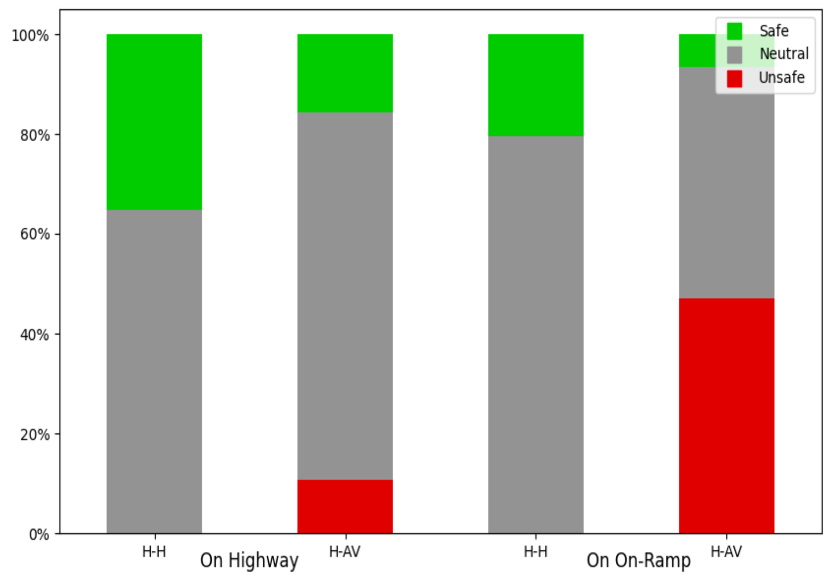

Figure 56: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

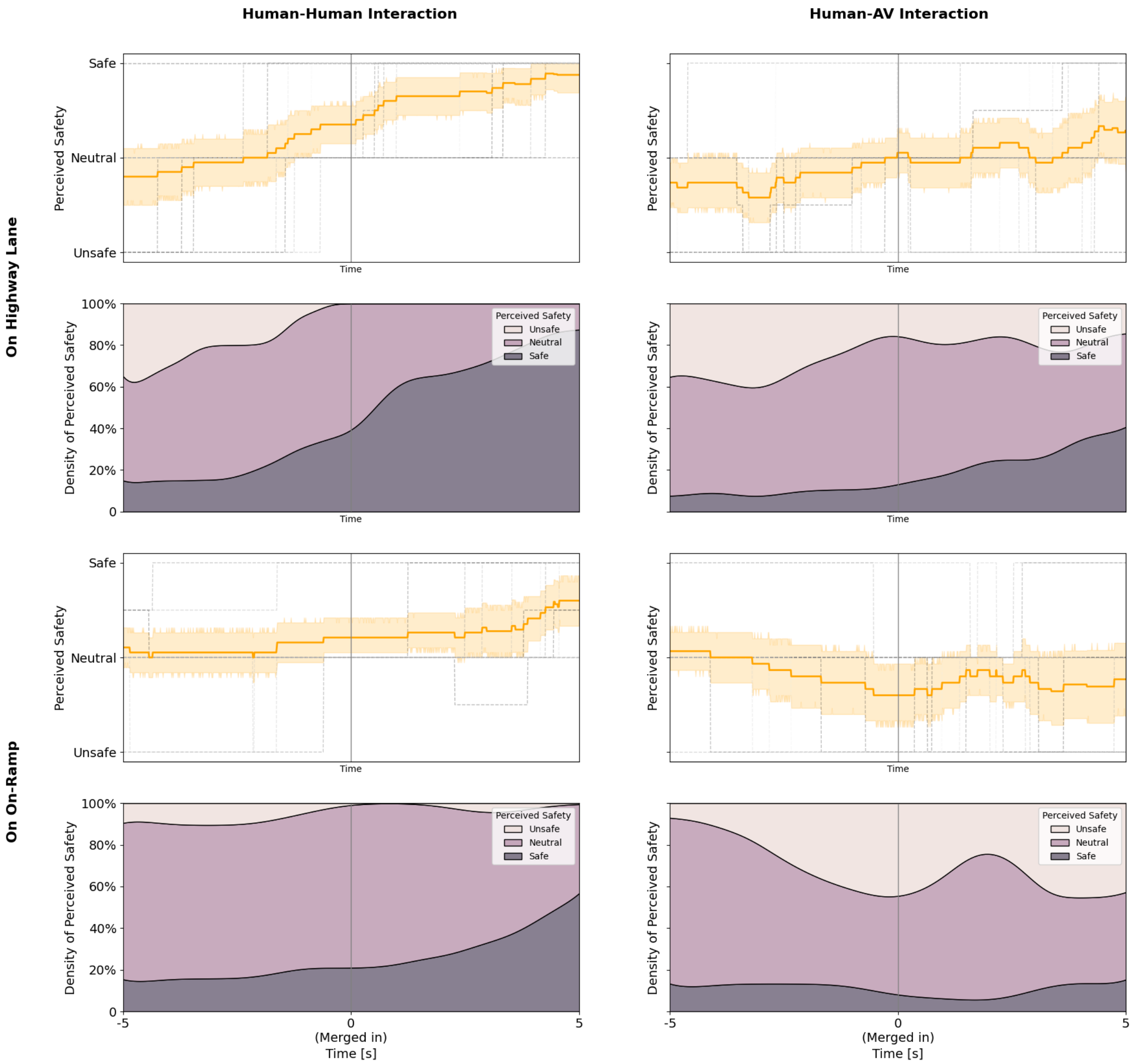

Figure 57: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

APPENDIX J

Appendix J shows the average lateral and longitudinal positions of the vehicles during experiment 10 divided by conditions Figure 58, the boxplots for the fixation frequency for experiment 10 Figure 59, the boxplots for the total duration of the fixation for experiment 10 Figure 60, the proportions of time of participants' Perceived Safety when the merging vehicle crossed the merging line for experiment 10 Figure 61 and the Perceived Safety during experiment 10 Figure 62.

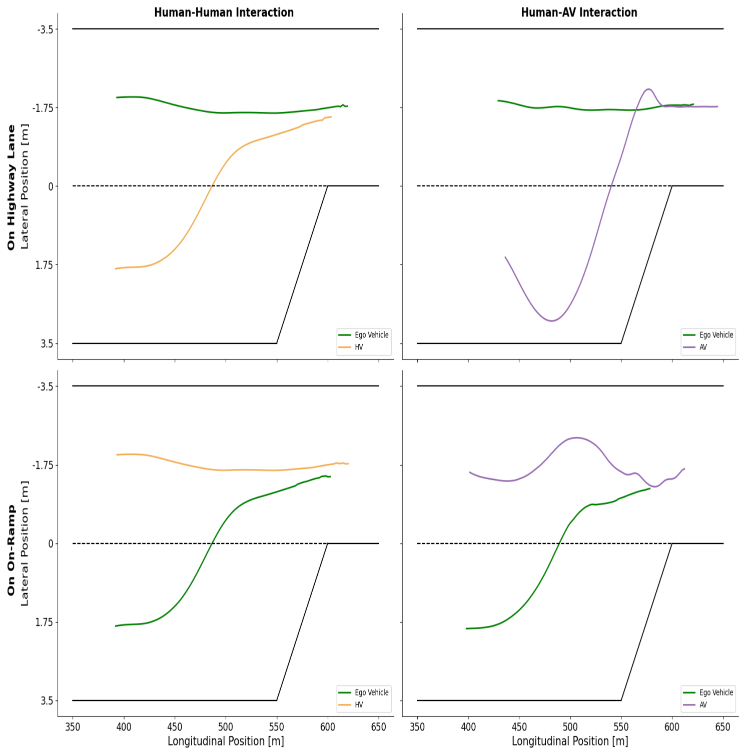

Figure 58: Average Lateral and longitudinal positions of the vehicles during experiment 1

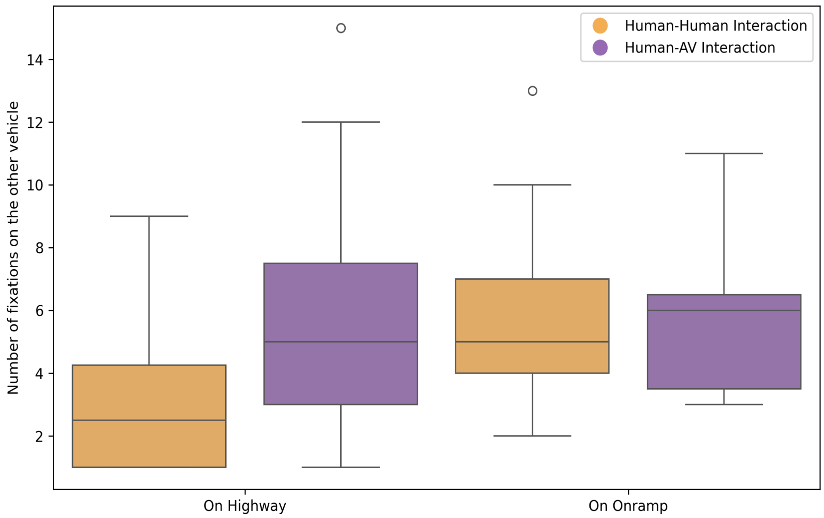

Figure 59: Numbers of time of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

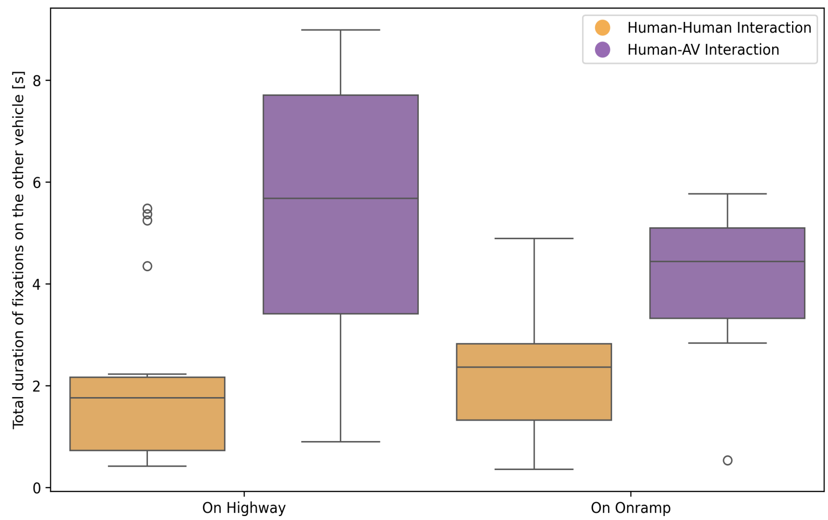

Figure 60: Total duration of fixations on the other vehicle as a function of experimental condition and interaction type for experiment 1

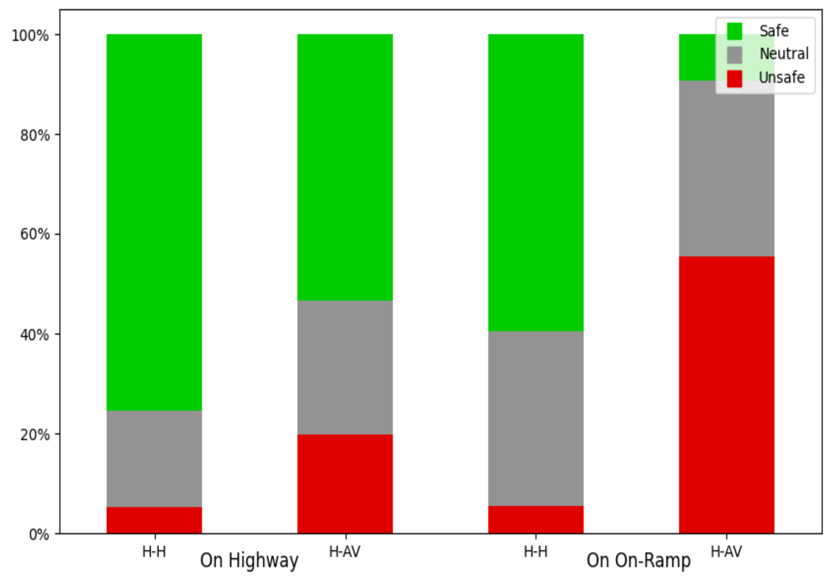

Figure 61: Proportion of perceived safety when the merging vehicle crossed the middle line for experiment 1

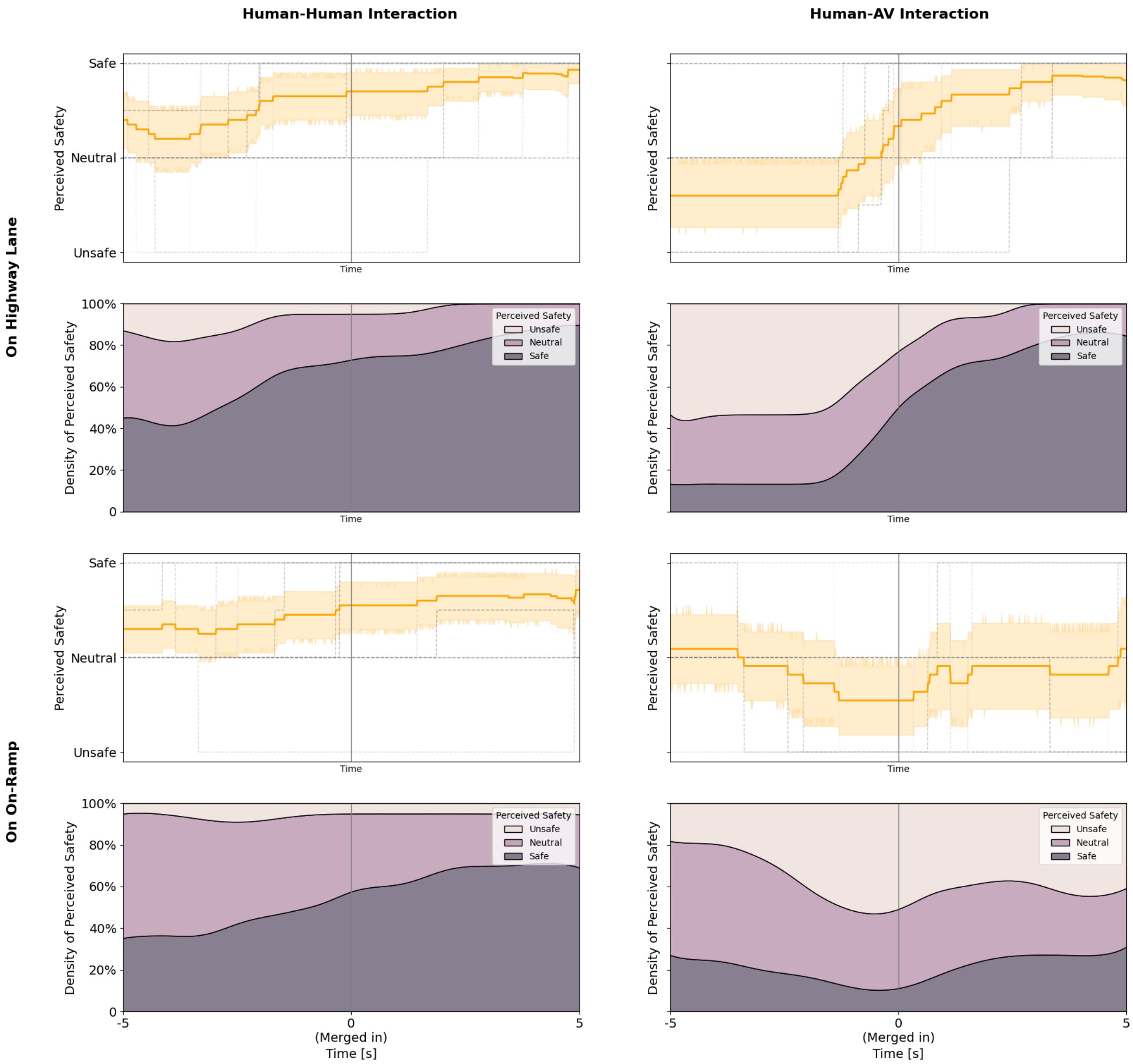

Figure 62: Perceived Safety and its density for experiment 1, divided by condition. The left side of the figure shows the trials with human interaction, and the right side depicts those with AV interaction. The top plots correspond to the "Highway" condition, while the bottom plots represent the "On-Ramp" condition. Each dotted line in these plots represents an individual trial, while the continuous orange line and surrounding shaded area illustrate the average and confidence interval, respectively, of the data for that condition. The grey dotted line marks the end of the merging manoeuvre. The three statuses of perceived safety: "Unsafe", "Neutral" and "Safe" are on the y-axis. Furthermore, all the datasets were shifted around the merging moment and shown between a fixed time window, to have a better understanding and comparison of the interaction at the same time.

Appendix K

Appendix K shows the average lateral and longitudinal positions of the vehicles during all experiments divided by conditions Figure 63, the pie-chart showing the percentage of participants that reported to have noticed differences in behaviour between the different coloured-coded cars,Figure 14, the boxplots for the total duration of the fixation for experiment 1 Figure 64 and the distribution of the answers on how large the noted differences were Figure 65.

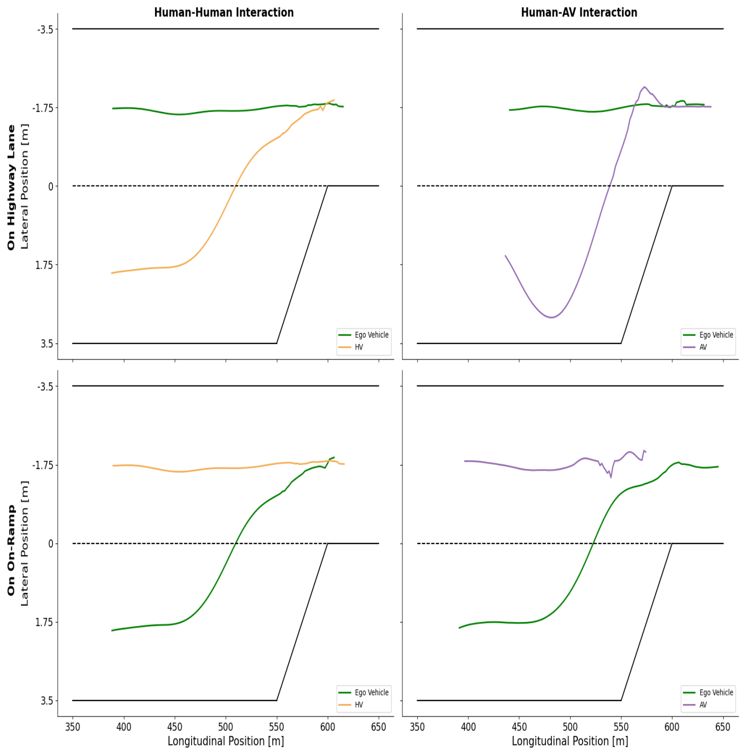

Figure 63: Average Lateral and longitudinal positions of the vehicles during experiment 1

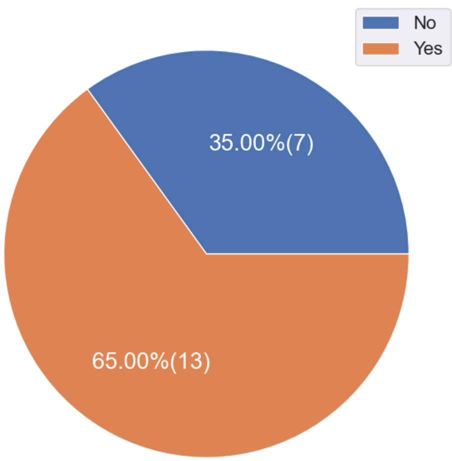

Figure 64: Percentage of people reporting if they noticed any difference in behaviour between the different coloured-coded cars. In blue is the percentage of participants who noticed a difference while in red is who didn't

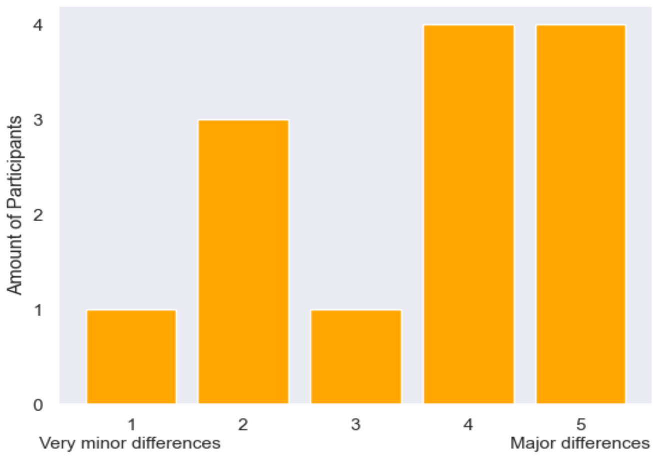

Figure 65: Distribution of the answers on how large the noted differences were. 13 answers since 7 participants didn't notice any difference